

Universidade Federal do Maranhão  
Centro de Ciências Exatas e Tecnologia  
Curso de Pós-Graduação em Engenharia de Eletricidade

---

*Métodos para sistemas CAD e CADx de  
nódulo pulmonar baseada em tomografia  
computadorizada usando análise de forma e  
textura*

---

Antonio Oseas de Carvalho Filho

São Luís  
2016

Universidade Federal do Maranhão  
Centro de Ciências Exatas e Tecnologia  
Curso de Pós-Graduação em Engenharia de Eletricidade

---

*Métodos para sistemas CAD e CADx de  
nódulo pulmonar baseada em tomografia  
computadorizada usando análise de forma e  
textura*

---

**Antonio Oseas de Carvalho Filho**

Tese apresentada ao Curso de  
Pós-Graduação em Engenharia de Eletricidade da UFMA  
como parte dos requisitos necessários para obtenção do  
grau de Doutor em Engenharia Elétrica.

Orientadores: **Prof. Dr. Aristófanés Corrêa Silva**  
**Prof. Dr. Anselmo Cardoso de Paiva**

**São Luís**  
**2016**

de Carvalho Filho, Antonio Oseas

Métodos para sistemas CAD e CADx de nódulo pulmonar baseada em tomografia computadorizada usando análise de forma e textura/ Antonio Oseas de Carvalho Filho. - São Luís, 2016.

144f.:il.

Co-Orientador: Anselmo Cardoso de Paiva.

Orientador: Aristófanês Corrêa Silva.

Tese (Doutorado) - Tese (Doutorado) - Programa de Pós-graduação em Engenharia de Eletricidade/ccet, Universidade Federal do Maranhão, Núcleo de Computação Aplicada - NCA, 2016.

1. Algoritmo Evolutivo. 2. Análise de Forma. 3. Câncer de Pulmão. 4. Imagens médicas. 5. Índice de Diversidade Filogenética. I. Paiva, Anselmo Cardoso de. II. Silva, Aristófanês Corrêa. III. Título.

**Métodos para sistemas CAD e CADx de nódulo  
pulmonar baseada em tomografia  
computadorizada usando análise de forma e  
textura**

**Antonio Oseas de Carvalho Filho**

Aprovado em //

**BANCA EXAMINADORA**

---

**Prof. Dr. Aristófanés Corrêa Silva**

Orientador

---

**Prof. Dr. Anselmo Cardoso de Paiva**

Coorientador

---

**Prof. Dr. Vicente Leonardo Paucar Casas**

Examinador Interno

---

**Prof. Dr. Allan Kardec Duailibe Barros Filho**

Examinador Interno

---

**Prof. Dr. Rodolfo Acatauassú Nunes**

Examinador Externo

---

**Prof<sup>a</sup>. Dr<sup>a</sup>. Fátima Nelsizeuma Sombra de Medeiros**

Examinador Externo

*”Por natureza, os homens são próximos; a educação é  
que os afasta”*

**Confúcio**

*A meus pais, irmão e irmã.*

# Agradecimentos

---

A Deus, por sua graça e todas as bênçãos e inspirações que a mim tem destinado.

Aos meus pais, Antonio Oseas de Carvalho e Francisca Raquel Rodrigues por todo o carinho e apoio imprescindíveis na minha vida. A meu irmão Ailton Antonio Rodrigues.

Ao meu orientador Prof. Aristófanés Silva pela confiança e credibilidade depositadas em mim desde o início desta jornada, afinal, foram quase 6 anos. Agradeço por seus ensinamentos, orientações eficazes e conselhos. Agradeço também por sua paciência, oportunidade de crescimento, compreensão nos momentos mais difíceis, por seu empenho em estar sempre presente quando solicitado. A este sempre terei respeito, amizade e admiração.

Ao meu coorientador Prof. Anselmo Paiva, por todas as oportunidades, por seus ensinamentos, pelos puxões de orelha, pelos valiosos conselhos. E como sempre disse, aqui não seria diferente: "o senhor é um avô para mim". Além de ser um *cara* espetacular.

Aos amigos, Wener e Stelmo, pelos conselhos, dicas e amizade, nesta rotina árdua, mas gratificante.

Aos meus companheiros de longa caminhada. Em especial aos que compõem o LABPAI. Muito obrigado por toda ajuda e companheirismo durante esta etapa. Aqui formou-se uma família.

Aos meus valiosos amigos de Picos, em especial a Alcilene, que foi uma irmã que Deus me concedeu, a Patricia Medyna, amiga e conselheira e a Márcio Alves por ler minha tese.

A todos, mais uma vez, os meus sinceros agradecimentos do fundo do coração.

## RESUMO

O câncer de pulmão é apontado como a principal causa de morte entre os pacientes com câncer. As altas taxas de mortes e registros de ocorrências desse câncer em todo o mundo demonstram a importância do desenvolvimento e investigação, a fim de produzir meios para a detecção e o diagnóstico precoce dessa doença. Devido ao exaustivo processo de análise, alternativas como ferramentas de cunho computacional que utilizam técnicas de processamento de imagens e do reconhecimento de padrões têm sido amplamente exploradas. Assim, para auxiliar o especialista na identificação e diagnóstico de nódulos, são desenvolvidos sistemas *Computer-Aided Detection (CAD)* e *Computer-Aided Diagnostic (CADx)*. Esta tese propõe o desenvolvimento de métodos para redução de falsos positivos em um sistema CAD e diagnóstico de nódulos em tomografia computadorizada. Os métodos propostos baseiam-se em técnicas de processamento de imagens e reconhecimento de padrões. Para tanto, foram adaptados e aplicados os conceitos da biologia no ramo do estudo da diversidade entre espécies, sendo esses os índices de diversidade filogenética, usados nesta tese como descritores de textura. Em outro aspecto, foram desenvolvidas e adaptadas técnicas capazes de mensurar propriedades de forma dos achados radiológicos. Seguindo, usou-se uma metodologia evolutiva genética para seleção dos melhores modelos de treinamento. E por fim, foi aplicada a máquina de vetor de suporte para realizar a classificação. Resultados promissores foram encontrados em teste com 833 exames divididos em 80% para treino e 20% para testes. Em linhas gerais, para os melhores resultados tem-se, nos métodos de redução de falsos positivos: uma acurácia de 99,57%, sensibilidade de 99,45%, especificidade de 99,61% e uma curva ROC de 0,992. Já nos resultados para a classificação quanto a taxa de malignidade e benignidade, obtiveram-se os seguintes valores: acurácia de 93,46%, sensibilidade de 92,95%, especificidade de 93,49% e uma curva ROC de 0,931.

**Palavras-Chave:** Imagens médicas, Câncer de Pulmão, Índice de Diversidade Filogenética, Análise de Forma, Algoritmo Evolutivo.



## ABSTRACT

Lung cancer has been identified as the leading cause of death among cancer patients worldwide. The high rates of deaths and instances of records of this type of cancer worldwide demonstrate the importance of the development and research in order to produce resources for the detection and early diagnosis of this disease. Because of the exhaustive analysis process, alternatives such as computational tools that use image processing techniques and pattern recognition have been widely explored. Therefore, to assist the expert in the identification and diagnosis of nodules, systems are developed *Computer-Aided Detection* (CAD) and *Computer-Aided Diagnostic* (CADx). This thesis proposes the development of methods that reduce false positives, and the diagnosis of volumes of interest in computed tomography. The proposed methods are based on image processing techniques and pattern recognition. For this, biology concepts have been adapted and applied to the study of the branch of the diversity of species; such concepts are the phylogenetic diversity indexes used in this thesis as texture descriptors. In another aspect, techniques that measure the properties of the shape of radiological findings have been developed and adapted. Subsequently, an evolutionary methodology is used  $\alpha$  for the selection of the best models for training. Finally, a support vector machine is applied to perform the classification. Promising results were found in the 833 tests that we performed; these tests were divided into 80% for training and 20% for testing. In general, for the best results, we have false positive reduction methods, an accuracy of 99.57%, sensitivity of 99.45%, specificity of 99.61%, and an ROC curve of 0.992. The results obtained for the classification of the degree of malignancy and benignity are: accuracy of 93.46%, sensitivity of 92.95%, specificity of 93.49%, and an ROC curve of 0.931.

**Keywords:** Lung Cancer, Medical images, Phylogenetic diversity index, Shape Analysis, Evolutionary Algorithm.

# Produção Científica

Tabela 1: Artigos submetidos e publicados da tese. Legenda: Publicado (PU), em revisão (ER) e submetido (SU).

Local	Artigo	Qualis	Status
Periódico	Lung-nodule classification based on computed tomography using taxonomic diversity indexes and an SVM	A2	PU
Periódico	Computer Aided Diagnosis System for Lung Nodules Based on Computed Tomography Using Shape Analysis, a Genetic Algorithm and SVM	A1	PU
Periódico	3D shape analysis to reduce false positives for lung nodule detection systems	A1	PU
Periódico	Computer Aided Diagnosis System of Lung Nodules Based on Computed Tomography using distance-based Phylogenetic Diversity index, Genetic Algorithm and SVM	A2	SU
Periódico	Computer Aided Diagnosis System of Lung Nodules Based on Computed Tomography using topology-based Phylogenetic Diversity index	A2	SU
Periódico	Differentiation patterns of benignity and malignancy in pulmonary nodules, based on Computed Tomography using minimum-path Phylogenetic diversity index, genetic algorithm and SVM	B1	SU

Tabela 2: Outras publicações e submissões. Legenda: Publicado (PU), em revisão (ER) e submetido (SU).

Local	Artigo	Qualis	Status
Periódico	<b>FILHO, A. O. de C.</b> ; 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: <a href="http://www.sciencedirect.com/science/article/pii/S09333657">http://www.sciencedirect.com/science/article/pii/S09333657</a>	A1	PU
Periódico	OLIVEIRA, F. S. S. de; <b>FILHO, A. O. de C.</b> ; SILVA, A. C.; PAIVA, A. C. 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, 2015. ISSN 0010-4825. Disponível em: <a href="http://www.sciencedirect.com/science/article/pii/S0010482514003357">http://www.sciencedirect.com/science/article/pii/S0010482514003357</a> .	A1	PU
Periódico	SANTOS, A. M.; <b>FILHO, A. O. de C.</b> ; SILVA, A. C.; PAIVA, A. C. de; NUNES, R. A.; GATTASS, M. Automatic detection of small lung nodules in 3d CT data using gaussian mixture models, tsallis entropy and SVM. Engineering Applications of Artificial Intelligence, v. 36, p. 27 39, 2014. ISSN 0952-1976. Disponível em: <a href="http://www.sciencedirect.com/science/article/pii/S0952197614001754">http://www.sciencedirect.com/science/article/pii/S0952197614001754</a>	A1	PU
Periódico	SILVA, J. N.; <b>FILHO, A. O. C.</b> ; SILVA, A. C.; PAIVA, A. Cardoso de; GATTASS, M. Automatic detection of masses in mammograms using quality threshold clustering, correlogram function, and svm. Journal of Digital Imaging, v. 28, n. 3, p. 323337, 2014. ISSN 1618-727X. Disponível em: <a href="http://dx.doi.org/10.1007/s10278-014-9739-3">http://dx.doi.org/10.1007/s10278-014-9739-3</a> .	A2	PU
Conferência	NETO, OTILIO PAULO S. ; <b>CARVALHO, OSEAS</b> ; SAMPAIO, WENER ; CORREA, ARISTOFANES ; PAIVA, ANSELMO . Automatic segmentation of masses in digital mammograms using particle swarm optimization and graph clustering. In: 2015 International Conference on Systems, Signals and Image Processing (IWSSIP), 2015, London. 2015 International Conference on Systems, Signals and Image Processing (IWSSIP), 2015. p. 109	B5	PU
Periódico	Classification of breast tissues into mass and non-mass by means of the Micro-genetic Algorithm, Phylogenetic Trees, LBP and SVM	-	ER
Periódico	Lung nodule classification using artificial crawlers, directional texture and support vector machine	A1	SU

# Lista de Tabelas

1	Artigos da tese. . . . .	9
2	Outras publicações e submissões. . . . .	10
2.1	Trabalhos relacionados para métodos de redução de FP. . . . .	32
2.2	Trabalhos relacionados para métodos de diagnóstico pulmonar. . . . .	37
3.1	Correspondência entre os termos da biologia e os métodos propostos. . . . .	55
3.2	Formas funcionais para os conjuntos em E2. . . . .	63
4.1	Redução de FP sem as regiões para medidas de textura. . . . .	102
4.2	Redução de FP com as regiões para medidas de textura. . . . .	102
4.3	Redução de FP sem as regiões para medidas de forma. . . . .	104
4.4	Redução de FP com as regiões para medidas de forma. . . . .	104
4.5	Comparação com métodos de redução de FP que usam descritores de forma. . . . .	110
4.6	Comparação com métodos de redução de FP que usam descritores de forma e textura. . . . .	111
5.1	Diagnóstico sem as regiões para medidas de textura . . . . .	117
5.2	Diagnóstico com as regiões para medidas de textura . . . . .	117
5.3	Diagnóstico sem as regiões para medidas de forma . . . . .	118
5.4	Diagnóstico com as regiões para medidas de forma . . . . .	119
5.5	Comparação com métodos que utilizam descritores de textura. . . . .	124
5.6	Comparação com métodos que utilizam descritores de forma e textura. . . . .	125

# Lista de Figuras

3.1	Exemplos de TC com nódulos benigno e maligno. . . . .	39
3.2	Aparência do nódulo pulmonar. . . . .	41
3.3	Exemplo de imagem com 3 níveis de cinza. . . . .	53
3.4	Exemplo de árvore filogenética. . . . .	53
3.5	Exemplo de uma árvore taxonômica. . . . .	54
3.6	Árvore filogenética na forma de dendrograma. . . . .	55
3.7	Funções geométricas. . . . .	62
3.8	Exemplo de cálculo do raio para formar o cilindro. . . . .	66
3.9	Exemplo de cilindros. . . . .	66
3.10	Exemplo do volume e da esfera calculados na região cilíndrica para: (A) VOI de nódulo e (B) VOI de não-nódulo. . . . .	67
3.11	Exemplos dos três cilindros gerados. . . . .	68
3.12	Exemplos de nódulos. . . . .	72
3.13	Exemplo em 2D dos descritores (A) $Md1$ e (B) $Md2$ . . . . .	73
3.14	Exemplos de círculos que serão geradas para cada ponto do esqueleto. Fonte (SAMPAIO <i>et al.</i> , 2015). . . . .	74
3.15	Área média do triângulo. . . . .	75
3.16	Exemplos de resultados para a TAR. . . . .	76
3.17	Exemplo de separação entre classes com MVS . . . . .	78
3.18	Ilustração da escolha do melhor modelo. Fonte (FILHO <i>et al.</i> , 2013). . . . .	82
3.19	Ilustração da curva ROC. . . . .	85
4.1	Método de redução de falsos positivos . . . . .	88
4.2	Exemplo de marcação em uma fatia TC. . . . .	89
4.3	Três exemplos VOIs: (A) nódulo, (B) não-nódulo, (C) outra estrutura. . . . .	91

4.4	Região intralesional. . . . .	92
4.5	Exemplo de histograma que apresenta duas classes bem definidas. . . . .	92
4.6	Regiões geradas com Otsu. . . . .	93
4.7	Iterações realizadas pelo algoritmo de Otsu. . . . .	95
4.8	Dedrograma gerado a partir da ROI mostrada na Figura 4.7. . . . .	96
4.9	Resultados para as curvas ROC dos métodos de redução de falsos positivos. . . . .	106
5.1	Método para o diagnóstico pulmonar. . . . .	114
5.2	Ilustração do resumo dos nódulos. Fonte (JABON <i>et al.</i> , 2009). . . . .	114
5.3	Exemplo do resumo do diagnóstico dos nódulos. Fonte (JABON <i>et al.</i> , 2009). . . . .	115
5.4	Resultados para as curvas ROC dos métodos de diagnóstico pulmonar. . . . .	121

# Lista de Abreviaturas e Siglas

ANN	<i>Artificial Neural Network</i>
CAD	<i>Computer-aided Detection</i>
CADx	<i>Computer-aided Diagnosis</i>
cm	Centímetro
DICOM	<i>Digital Imaging Communications in Medicine</i>
FN	Falso Negativo
FP	Falso Positivo
GLCM	<i>Gray-Level Cooccurrence Matrix</i>
KNN	<i>k-Nearest Neighbor</i>
LBP	<i>Local Binary Pattern</i>
LDA	<i>Linear Discriminant Analysis</i>
LDCT	<i>Low Dose Computed Tomography</i>
LIDC	<i>Lung Image Database</i>
LIDC-IDRI	<i>Lung Image Database Consortium Image Collection</i>
ME $\alpha$	Metodologia de Evolução $\alpha$
MVS	Máquina de Vetores de Suporte
mm	Milímetros
NCR	National College of Radiography
NEMA	National Electrical Manufacturers Association
NCI	<i>National Cancer Institute</i>
NPS	Nódulo Pulmonar Solitário
OMS	Organização Mundial de Saúde
PCA	<i>Principle Component Analysis</i>
RBF	<i>Radial Base Function</i>
RF	<i>Random Forest</i>
ROC	<i>Receiver Operationg Characteristic</i>
ROI	<i>Region of Interest</i>

SIFT	<i>Scale-Invariant Feature Transform</i>
TC	Tomografia Computadorizada
UH	Unidade de Hounsfield
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo
VOI	<i>Volume of Interest</i>



# Sumário

<b>1</b>	<b>Introdução</b>	<b>19</b>
1.1	Definição do Problema . . . . .	22
1.2	Objetivos e Solução . . . . .	23
1.3	Contribuições . . . . .	24
1.4	Organização da tese . . . . .	25
<b>2</b>	<b>Trabalhos Relacionados</b>	<b>27</b>
2.1	Redução de falsos positivos . . . . .	27
2.2	Diagnóstico . . . . .	33
<b>3</b>	<b>Fundamentação Teórica</b>	<b>38</b>
3.1	Fisiologia, Imagem e Patologia Pulmonar . . . . .	38
3.1.1	Nódulo Pulmonar . . . . .	39
3.2	Tomografia Computadorizada . . . . .	42
3.3	Arquivo DICOM . . . . .	43
3.3.1	Padrão DICOM . . . . .	44
3.4	Base LIDC-IDRI . . . . .	44
3.5	Sistemas Computacionais de Auxílio à Detecção e ao Diagnóstico	46
3.6	Processamento Digital de Imagens . . . . .	48
3.6.1	Segmentação . . . . .	49
3.6.2	Extração de Características . . . . .	51
3.7	Análise por Textura . . . . .	51
3.7.1	Índices de Diversidade e Diversidade Filogenética . . . . .	52
3.8	Análise por Forma . . . . .	61
3.8.1	Diagramas de Forma 2D . . . . .	62
3.8.2	Abordagem Cilíndrica 3D . . . . .	65

3.8.3	Funcionais de Minkowski . . . . .	68
3.8.4	Medidas de Distância . . . . .	71
3.8.5	Medidas Baseadas no Esqueleto . . . . .	73
3.8.6	Medidas Baseadas em Triangulação . . . . .	74
3.8.7	Medidas Baseadas nos Diâmetros de Feret . . . . .	76
3.9	Reconhecimento de Padrões . . . . .	77
3.9.1	Máquina de Vetor de Suporte . . . . .	78
3.9.2	Algoritmo evolutivo . . . . .	80
3.9.3	Análise Discriminante <i>Stepwise</i> . . . . .	82
3.9.4	Validação dos resultados . . . . .	83
3.10	Considerações Finais . . . . .	86
<b>4</b>	<b>Redução de Falsos Positivos</b>	<b>87</b>
4.1	Método . . . . .	87
4.1.1	Aquisição de Imagens . . . . .	87
4.1.2	Regiões Internas . . . . .	91
4.1.3	Extração de Características . . . . .	93
4.1.4	Reconhecimento de Padrões . . . . .	97
4.1.5	Validação dos resultados . . . . .	99
4.2	Resultados . . . . .	99
4.2.1	Resultados para os Descritores de Textura . . . . .	101
4.2.2	Resultados para os Descritores de Forma . . . . .	103
4.2.3	Resultados com todos os descritores e seleção de características . . . . .	105
4.3	Discussão . . . . .	106
4.4	Comparação com outros Trabalhos . . . . .	109
4.5	Considerações Finais . . . . .	112
<b>5</b>	<b>Diagnóstico de nódulos pulmonares</b>	<b>113</b>
5.1	Método . . . . .	113
5.1.1	Aquisição de Imagens . . . . .	113
5.2	Resultados . . . . .	116
5.2.1	Resultados para os Descritores de Textura . . . . .	116
5.2.2	Resultados para os Descritores de Forma . . . . .	118

---

5.2.3	Resultados com todos os descritores e seleção de características . . . . .	120
5.3	Discussão . . . . .	121
5.4	Comparação com outros Trabalhos . . . . .	123
5.5	Considerações Finais . . . . .	126
<b>6</b>	<b>Conclusão</b> . . . . .	<b>127</b>
6.1	Trabalhos Futuros . . . . .	130
	<b>Referências</b> . . . . .	<b>131</b>

## CAPÍTULO 1

# Introdução

---

Câncer é uma denominação genérica para um conjunto de mais de 100 doenças que possuem em comum o crescimento desordenado de células que invadem os tecidos e órgãos. As causas do câncer são variadas, podendo ser externas ou internas ao organismo, estando ambas inter-relacionadas. As causas externas relacionam-se ao meio ambiente e aos hábitos ou costumes próprios de um ambiente social e cultural. As causas internas são, na maioria das vezes, geneticamente pré-determinadas e estão ligadas à capacidade do organismo de se defender das agressões externas (SPROUFFSKE *et al.*, 2013).

O câncer de pulmão vem se tornando uma das principais causas de morte no mundo, visto seu aumento anual de 1,8 milhões de casos. Isto representa um aumento de 13% no total de casos de câncer (CHEN *et al.*, 2011; PATIL *et al.*, 2014; STEWART; WILD, 2014). De acordo com Chen *et al.* (2011), o tabagismo, o envelhecimento da sociedade, a urbanização, a poluição, o estilo de vida ruim e a industrialização são os principais fatores relacionados aos casos de câncer de pulmão. Entretanto, o fator crucial de ocorrência desse tipo de câncer é marcado pela grande exposição ao tabagismo acometido no passado. Na maioria das populações, os casos de câncer de pulmão que possuem o tabaco relacionado representam 80% ou mais dos casos desse tipo de câncer (CHEN *et al.*, 2012). Comparados com os não fumantes, os tabagistas têm cerca de 20 a 30 vezes mais risco de desenvolver o câncer. Em geral, as taxas de incidência do câncer de pulmão em um determinado país refletem diretamente o consumo do cigarro. A sobrevida de cinco anos varia entre 13 e 21% em países desenvolvidos de 7 a 10%

nos países subdesenvolvidos (CHEN *et al.*, 2011).

Apesar de sua incidência apresentar uma diminuição no Brasil por conta da diminuição no número de fumantes, o câncer pulmonar ainda apresenta um elevado número de casos. Essa diminuição se deu, principalmente, devido às campanhas de prevenção contra o tabagismo. É importante ressaltar que o número de casos em pessoas não fumantes está aumentando, isto é, não apenas os fumantes passivos estão apresentando a doença. Acredita-se que fatores oriundos da globalização, como poluição, urbanização, entre outros, estejam diretamente relacionados com essa ocorrência.

No Brasil, as estimativas para 2014/2015 eram em média de 27.330 novos casos de câncer de pulmão (10.930 em homens e 16.400 em mulheres). O câncer de pulmão é uma das principais causas de morte por câncer entre homens e mulheres. Pois, devido alta taxa de incidência do câncer pulmonar, esse acaba se tornando o mais letal. A cada ano, mais pessoas morrem de câncer de pulmão do que de câncer colorretal, câncer de mama e câncer de próstata combinados. O câncer de pulmão ocorre principalmente em pessoas mais velhas. Cerca de 2 em cada 3 pessoas com diagnóstico de câncer de pulmão têm mais de 65 anos, e menos de 2% dos casos são diagnosticados em pessoas com menos de 45 anos. A idade média no momento do diagnóstico é de 70 anos. Em geral, a chance de um homem desenvolver câncer de pulmão em sua vida é de cerca de 1 em 13, enquanto para uma mulher, esse risco é de cerca de 1 em 16. Esses números incluem tanto os fumantes, como os não fumantes. Para os fumantes, o risco é muito maior, enquanto que para os não fumantes, esse risco é menor (INCA, 2015).

Um nódulo pulmonar é caracterizado por uma opacidade arredondada no pulmão com até 3 cm de diâmetro, envolto 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). As lesões geralmente são subdivididas de acordo com o seu tamanho, por isso, lesões com diâmetro menores que 8-10 mm apresentam menor probabilidade de malignidade e têm diferentes recomendações para a análise, em comparação com nódulos pulmonares maiores. O tamanho de um nódulo pulmonar é um fator relevante para auxiliar na diferenciação entre benignos e malignos. Como regra geral, nódulos maiores apresentam maior

probabilidade de câncer (MOSMANN *et al.*, 2016), a qual varia muito com os tamanhos dos nódulos. Cerca de 80% dos nódulos benignos possuem menos de 2 cm de diâmetro. No entanto, 15% dos nódulos malignos são menores que 1 cm e aproximadamente 42% são menores que 2 cm (ERASMUS *et al.*, 2000).

O diagnóstico e o tratamento precoce do câncer pulmonar aumentam as chances de sobrevivência do paciente em cerca de 90% (LEDERLIN *et al.*, 2013). A maneira mais eficaz para se defender do câncer de pulmão é o diagnóstico precoce e o tratamento. Se detectada precocemente, a taxa de sobrevivência pós-diagnóstico aumenta cerca de 90% (LEDERLIN *et al.*, 2013). Um dos métodos mais eficazes para o diagnóstico é o exame por imagem. A Tomografia Computadorizada (TC) é um exame acessível aos pacientes, por conta do seu baixo custo, se comparado a outros exames por imagem, e que fornece imagens de boa qualidade, sendo essa utilizada na análise de vários tipos de lesões, incluindo lesões pulmonares.

Devido a boa resolução do exame de TC, muitas imagens são produzidas no momento do exame. Assim, os especialistas têm que analisar imagens de TC reconhecendo os nódulos pulmonares menores, que possuem um grau de detecção mais elevado, isto é, levando em consideração a área na qual o nódulo está localizado e também a possibilidade de conter estruturas conexas a ele (como vasos sanguíneos). Esse processo de análise da TC para retirada dos dados mais relevantes, é de extrema complexidade, uma vez que é um processo repetitivo e exaustivo, isto é, qualquer desvio na atenção do especialista pode ocasionar algum equívoco na análise, especialmente quando há outras anomalias na mesma imagem de TC. Dessa forma, a vulnerabilidade a erros nesse tipo de análise torna-se frequente.

Quanto mais informações disponibilizadas ao especialista, mais preciso será o diagnóstico. Além disso, a detecção nos estágios iniciais do câncer pulmonar permite ao paciente uma intervenção terapêutica mais rápida e, conseqüentemente, um prognóstico mais benéfico (SONE *et al.*, 1998). Com esse intuito, ferramentas computacionais que utilizam processamento de imagens digitais e reconhecimento de padrões para desenvolver sistemas *Computer-Aided Detection* (CAD) e *Computer-Aided Diagnostic* (CADx) têm sido pesquisadas e estudadas, proporcionando ao especialista uma segunda opinião (YE *et al.*, 2009a).

Segundo Ye *et al.* (2009b), os sistemas CAD têm o potencial de analisar a morfologia dos nódulos e melhorar o fluxo de trabalho durante a avaliação dos estudos de segmentação. Além da identificação, existem também CAD desenvolvidos para o diagnóstico (malignos e benignos), denominados CADx.

O desenvolvimento de métodos automáticos para redução de falsos positivos e para o diagnóstico estão sendo amplamente exploradas em todo o mundo. Normalmente, os processos envolvidos nos sistemas CAD e CADx são compostos por quatro fases: 1) a aquisição de imagem; 2) extração/segmentação dos volumes de interesse ou nódulos para análise; 3) extração de características; e 4) classificação. Nesse contexto, tem-se como objetivos centrais desenvolver soluções que através da exploração das análises de forma e textura, sejam capazes de mensurar sobre a diferenciação entre classes, isto é, entre nódulo e não-nódulo, quando se tratar de redução de falsos positivos; e em maligno e benigno, quando se remeter ao diagnóstico.

## 1.1 Definição do Problema

Os avanços alcançados com as ferramentas de detecção e diagnóstico representam a capacidade que possui a análise de exames por imagens. Todavia, problemas comumente encontrados em ferramentas dessa natureza necessitam ser explorados. Tais problemas relacionam-se às taxas de sensibilidade e especificidade. As ferramentas que possuem finalidade apenas de detecção ainda possuem como dificuldade altas taxas de falsos positivos. Contudo, problemas rotineiros presentes nos exames de tomografia computadorizada, como: baixo contraste, localização do nódulo com anatomia complexa e nódulos conectados a vasos sanguíneos contribuem para possíveis erros no diagnóstico. Além disso, o processo repetitivo de análise de altas quantidades de imagens pelo especialista tornam o processo de análise exaustivo e propício a equívocos. Assim, o desenvolvimento de técnicas de extração de características de forma e textura podem auxiliar os especialistas na realização de diagnósticos mais precisos.

Dessa maneira, o problema abordado nesta tese consiste em apresentar métodos desenvolvidas que permitam, com bons níveis de acerto, realizar as tarefas de classificação (redução de falsos positivos) e diagnóstico pulmonar

(maligno e benigno) de Volumes de Interesse (VOI<sup>1</sup>) em imagens de TC. Para tanto, faz-se necessário abordar:

- A criação de técnicas que viabilizem análises da forma e da textura para caracterização de volumes de interesse em imagens de TC;
- Adaptar técnicas de processamento de imagens para tratamento genérico das imagens de TC, para delimitação de regiões;
- Desenvolver novas estratégias para definições de regiões internas aos objetos de estudo, que viabilizem uma análise com maior número de informações;
- Esquema de representação da imagem, por meio do qual se torne viável adaptar o conceito de índices de diversidade filogenéticos para a extração de características que mensuram a textura;
- Aplicar técnicas de otimização para criação de modelos de treinamento baseados em algoritmos evolutivos ; e
- Aplicar técnica de aprendizado de máquina para o tratamento genérico dos padrões.

## 1.2 Objetivos e Solução

O objetivo geral desta tese é apresentar o desenvolvimento de métodos para redução de falsos positivos e o diagnóstico (maligno e benigno) de volumes de interesse em imagens de TC, que proporcionem uma diminuição no número de falsos positivos e um aumento na sensibilidade de sistemas CAD e CADx.

As soluções propostas por esta tese para cumprir o exposto no objetivo geral, dar-se-á através da criação de novos descritores de forma, na utilização e adaptação da: técnica de segmentação desenvolvida por *Otsu* (YANG *et al.*, 2012), análise de textura usando índices de diversidade filogenéticos (PIENKOWSKI *et al.*, 1998; MAGURRAN, 2004; SILVA; BATALHA, 2006; VANDAMME *et al.*, 1996), análise de forma usando diagramas de forma (RIVOLLIER *et al.*, 2010b, 2010a), funcionais de Minkowski (HADWIGER, 1957), medidas de proporção,

---

<sup>1</sup>Do inglês Volume of Interest.



medidas de triangulação, medidas do esqueleto, medidas de Feret, algoritmos genéticos (FILHO *et al.*, 2013) e reconhecimento de padrões usando Máquina de Vetores de Suporte (VAPNIK; VAPNIK, 1998).

De maneira específica, pretende-se:

- Estudar, adaptar e aplicar algoritmos de segmentação para delimitação automática de regiões que possibilitem um estudo mais detalhado, baseado na análise de forma e textura;
- Criar meios eficientes para delimitação de regiões internas aos objetos de estudo, com a finalidade de evidenciar características fundamentais no discernimento entre classes;
- Estudar, adaptar e aplicar índices de diversidade filogenética para caracterização da textura dos objetos de estudo;
- Estudar, adaptar e desenvolver descritores de forma para descrição dos objetos de estudo;
- Estudar e adaptar algoritmos genéticos com o objetivo de otimizar o processo de classificação selecionando os indivíduos que melhor representem uma população;
- Estruturar os métodos para discriminar padrões nas etapas de redução de falsos positivos e diagnóstico nas imagens de tomografia computadorizada; e
- Avaliar os métodos propostos através de experimentos, utilizando uma base de imagens pública de tomografia computadorizada.

### 1.3 Contribuições

De uma forma mais abrangente, as principais contribuições desta tese são:

- No desenvolvimento de métodos para redução de falsos positivos e diagnóstico, usando análises de forma e textura;
- No desenvolvimento e adaptação de descritores, especificamente na:

- Adaptação de:
  - \* Índices de diversidade filogenética baseado na: riqueza de espécies, distância entre pares de espécies, topologia e em caminho mínimo; e
  - \* Diagramas de forma e funcionais de Minkowski.
- Criação da:
  - \* Abordagem com cilindros; e
  - \* Medidas de: proporção, distância, esqueleto, triangulação e diâmetros de Feret.
- Construção de estratégias para delimitação automática de regiões internas ao volume de interesse; e
- Construção de uma representação da imagem para uso dos índices de diversidade filogenética.

## 1.4 Organização da tese

Os demais capítulos desta tese foram organizados em:

- O Capítulo 2 apresenta um resumo dos trabalhos relacionados às soluções proposta por este trabalho, isto é, trabalhos que tratam do problema de redução de falsos positivos e diagnóstico em maligno e benigno;
- O Capítulo 3 trata da fundamentação teórica necessária para a construção da pesquisa proposta. São abordados os temas referentes ao câncer de pulmão, tomografias computadorizada do tórax, base de imagens utilizada, segmentação baseada no algoritmo de Otsu, índices de diversidade filogenéticos, descritores de forma, algoritmo genético, reconhecimento de padrões e máquinas de vetores de suporte;
- Os Capítulos 4 e 5 descrevem todas as etapas dos métodos propostos por esta tese, assim como os resultados discussões e comparação com outros trabalhos. Sendo que, o Capítulo 4 disserta sobre os métodos empregados

---

para redução de falsos positivos e o Capítulo 5 trata sobre as soluções propostas para o diagnóstico do nódulo pulmonar; e

- Por fim, o Capítulo 6 apresenta as conclusões e propostas de trabalhos futuros.

# Trabalhos Relacionados

---

Neste capítulo, são apresentados resumos dos trabalhos relacionados aos objetivos propostos. Vários estudos são frequentemente realizados com o objetivo de reduzir as taxas de falsos positivos e de elevar os acertos em sistemas que diferenciam entre padrões de malignidade e benignidade. Nesses termos, as seções seguintes enumeram os trabalhos relacionados, os quais estão divididos em trabalhos que tratam de redução de falsos positivos e trabalhos que apresentam métodos de diagnóstico, ambos voltados para nódulos pulmonares, objeto de estudo desta tese.

## 2.1 Redução de falsos positivos

Após a segmentação, VOIs incorretos podem ser gerados, conhecidos como falsos positivos. Um desafio importante é reduzir os falsos positivos sem perder os VOIs reais, isto é, os nódulos. A determinação de quais medidas de textura ou forma devem ser utilizadas em um sistema CAD é uma grande dificuldade em qualquer sistema. Nesta seção, são apresentados alguns trabalhos que estão intimamente relacionados com os métodos propostos neste trabalho.

Em Dehmeshki *et al.* (2006) é apresentado um método chamado de casamento de padrões de algoritmos genéticos para detecção automática de nódulos pulmonares. O cálculo da função de adaptação é baseado na forma geométrica do *voxel* e, em seguida, combinado com a distribuição global da intensidade do nódulo. Os autores relatam uma taxa de 14 falsos positivos

por exame. A metodologia foi avaliada com um conjunto de 70 imagens de TC contendo 178 nódulos, conseguindo uma sensibilidade de 85%.

Em Hardie *et al.* (2008) é apresentado um mecanismo de um sistema CAD para detecção de nódulos pulmonares. Os testes foram realizados sob um conjunto de dados que continham 167 radiografias de tórax, com 181 nódulos pulmonares. O sistema utilizava um algoritmo de limiar adaptativo baseado na distância entre os pontos, aplicado na etapa de segmentação dos nódulos. Seguindo, eram extraídas medidas baseadas na forma, intensidade e gradiente, para caracterização dos candidatos a nódulos. E na etapa de classificação, foi usado *Fisher discriminant linear*, alcançando uma sensibilidade de 78,1% e uma taxa de 4 falsos positivos (FP) por imagem.

O método apresentado por Horsthemke *et al.* (2009) propõe a definição de uma região de interesse de borda através da combinação de contornos anotados por radiologistas. Em seguida, é adaptado um método de indexação de gradiente radial para uso dentro das regiões da imagem e, por último, são feitas várias classificações do conjunto de anotações dos radiologistas em relação às características de forma, como: espiculação, lobulação e esfericidade. Os resultados foram validados usando a base LIDC e alcançaram uma taxa de acerto de 70%.

O trabalho proposto por Liu *et al.* (2009) apresenta uma metodologia dividida em etapas, incluindo a segmentação pulmonar, realce dos candidatos para nódulos pulmonares, extração de características e a classificação, com a utilização de Máquina de Vetores de Suporte (MVS) com função *Radial Basis Function* (RBF). Foi usada, para validar os resultados uma base de imagens contendo informações de 32 pacientes com nódulos pulmonares. Por fim, encontrou-se uma sensibilidade de 93,75%, especificidade de 87,6%, acurácia de 87,8% e uma taxa de FP de 4,6 por exame.

Para realizar a tarefa de classificação de nódulos pulmonares, a metodologia proposta em Jing *et al.* (2010) apresenta uma abordagem que combina *rule-based* e MVS. Primeiramente, são extraídas medidas das *Region of Interest* (ROI) candidatas baseadas na forma, facilitando a retirada de alguns vasos sanguíneos. Em seguida, são extraídas novas medidas dos candidatos restantes baseadas em textura e, por fim, estas serão as entradas do classificador

MVS. A metodologia foi aplicada em um conjunto de testes contendo 50 fatias, com 50 nódulos e 204 não-nódulos, alcançando uma especificidade de 92% e uma acurácia de 84,39%.

Em Lee *et al.* (2010), tem-se uma classificação auxiliada por um método baseado em *cluster*. Os experimentos são realizados utilizando exames de 32 pacientes, incluindo 5721 imagens, com os nódulos previamente marcados pelos especialistas, atingindo como melhor resultado uma sensibilidade de 97,33%, especificidade de 97,11% e uma área sob a curva *Receiver Operating Characteristic* (ROC) de 0,978.

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 ROC. Os resultados alcançados demonstram que a efetividade do método proposto é de 74,1% e uma curva ROC de 0.69.

A combinação de técnicas apresentadas por Farag *et al.* (2011) é o que define este trabalho. As técnicas são: Transformada *Scale Invariant Feature Transform* (SIFT), Local Binary Pattern (LBP), análise de componentes principais (PCA) e *Linear Discriminant Analysis* (LDA). As primeiras combinações foram PCA-SIFT e PCA-LBP, que apresentaram sensibilidade de 85%. A combinação de LDA-SIFT e LDA-LBP alcançaram a mesma sensibilidade.

Em Orozco *et al.* (2012) 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 usada uma MVS com RBF. A metodologia foi validada em 75 exames, atingindo resultados nas proporções de 10 falsos negativos (FN) e 2 FP, sensibilidade e especificidade de 96,15% e 52,17%, respectivamente, e uma acurácia de 82,66%.

O método apresentado em Al-Absi *et al.* (2012) mostra uma combinação de técnicas de aprendizado de máquina que compõem um sistema CAD. Esse sistema consiste em três grandes fases: (1) a extração de características, (2) a fase

de seleção de características e (3) a fase de classificação. O método é aplicado em um conjunto de imagens que possuem 154 regiões de nódulos e 92 regiões de não-nódulos, alcançando uma acurácia de 96,58%.

Segundo Tartar *et al.* (2013), a detecção precoce de nódulos pulmonares é extremamente importante para o diagnóstico e o tratamento de câncer de pulmão. Nesse âmbito, os autores desenvolveram uma nova abordagem para classificação de nódulos pulmonares em imagens de TC. Quatro métodos diferentes são introduzidos para o sistema proposto. A classificação é feita por diversos classificadores. O método proposto alcançou 90,7% de acurácia, 89,6% de sensibilidade e 87,5% de especificidade.

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 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, classificadas usando MVS. Os resultados alcançados apontam uma sensibilidade de 85,91%, uma especificidade de 97,70% e uma acurácia de 97,55%.

A metodologia proposta em Sivakumar e Chandrasekar (2013) apresenta um método de classificação baseado em descritores híbridos. A etapa de extração de características foi realizada utilizando a Análise de Componentes Principais bidimensional (2D-PCA) e o processamento de imagem morfológica. Além disso, na fase de classificação, foram utilizados métodos para selecionar os descritores mais significativos e, finalmente, Redes Neurais Artificiais, *Random Forest* (RF), *Bagging* e *AdaBoost* foram usados para classificar os candidatos. Os resultados alcançaram 90,7% de acurácia, 89,6% de sensibilidade e 87,5% 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 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%.

No método desenvolvido por Eskandarian e Bagherzadeh (2015), é

proposto um CAD de nódulos pulmonares utilizando algoritmo genético em imagens de TC e MVS para etapa de classificação. A metodologia foi validada utilizando a base LIDC com uma amostra de 147 pacientes. O resultado do sistema proposto obteve 90% de sensibilidade e uma taxa de 3,9 de falsos positivo.

Em Setio *et al.* (2016), os autores apresentam um sistema CAD utilizando multi-view redes convolucionais (ConvNets). Os testes foram realizados pela combinação de três tipos de candidatos a nódulo, que são: os sólidos, sub-sólidos e grandes nódulos. Esses testes foram realizados a partir da base LIDC com uma amostra de 888 nódulos. O método proposto compreende um múltiplo fluxo de ConvNets 2-D e o resultado obteve sensibilidade de 90,1%.

No CAD desenvolvido por Teramoto *et al.* (2016) é apresentada uma redução de FP através de análise por forma e redes neurais convolucionais (CNN). Na imagem de TC, uma região é detectada pela primeira vez usando um filtro de contorno ativo. Posteriormente, as regiões de alta absorção detectadas pelas imagens Positron emission tomography (PET) são mescladas com as regiões detectadas pelas imagens de TC. Os candidatos a nódulo são eliminados usando um conjunto de características extraídas a partir da análise de forma e outra pela CNN. Por fim, a classificação é feita em duas etapas, uma baseada em regras e outra baseada em MVS. Foram usadas 104 imagens para validação do método. Depois de executar o método proposto, a sensibilidade foi de 90,1% com 4,9 FPs/exame.

Esses são exemplos de sistemas que foram desenvolvidos para a detecção/redução de falsos positivos de VOIs de nódulo pulmonar. A Tabela 2.1 resume as abordagens. Três pontos importantes comuns desses sistemas são um número elevado de falsos positivos, uma elevada taxa de falsos negativos e um pequeno número de casos analisados. Além disso, os resultados de sensibilidade e especificidade foram desequilibrados. Em nossos métodos, tentar-se-á explorar essas deficiências com o intuito de melhorar as métricas que avaliam o desempenho, como acurácia, sensibilidade e especificidade.



Tabela 2.1: Resumo dos trabalhos relacionados que descrevem métodos de redução de falsos positivos em VOIs de nódulo pulmonar.

Trabalho	Técnicas	Base	Amostra
(DEHMESHKI <i>et al.</i> , 2006)	Algoritmo genético baseado em <i>template matching</i> , medidas geométricas e distribuição de intensidade.	Proprietário	70
(HARDIE <i>et al.</i> , 2008)	Algoritmo adaptativo baseado em limiar, atributos de forma, intensidade, e gradiente, <i>Fisher Discriminant Linear classifier</i> .	JSRT	167
(HORSTHEMKE <i>et al.</i> , 2009)	<i>Radial gradient indexing</i> , medidas de forma baseadas em: espiculação, lobulação e esfericidade.	LIDC	85
(LIU <i>et al.</i> , 2009)	Método do limiar, medidas de textura e MVS.	NSRCT-LUNG/LIDC	60
(JING <i>et al.</i> , 2010)	Características baseadas na forma e na textura, combinadas com <i>rule-based</i> e MVS.	3AHG	50
(LEE <i>et al.</i> , 2010)	<i>Random forest</i> e agrupamento.	LIDC	32
(ZINOVEV <i>et al.</i> , 2011)	Descritores baseados na textura e na forma, árvores de decisão baseado em confiança.	LIDC	32
(FARAG <i>et al.</i> , 2011)	<i>Scale-Invariant Feature Transform</i> , LBP, PCA e LDA.	ELCAP	70
(OROZCO <i>et al.</i> , 2012)	Descritores baseados em textura, histograma e MVS.	NBIA/ELCAP	75
(AL-ABSI <i>et al.</i> , 2012)	Aprendizado de máquina, Transformada de <i>Wavelet</i> , k-Nearest Neighbor (KNN), K-means.	JSRT	246
(TARTAR <i>et al.</i> , 2013)	2D-PCA, redundância mínima, relevância máxima, Artificial Neural Network (ANN), RF, Bagging e AdaBoost.	University of Istanbul	63
(FILHO <i>et al.</i> , 2013)	<i>Quality threshold clustering</i> , algoritmo genético, índice de diversidade e MVS.	LIDC-IDRI	833
(SIVAKUMAR; CHANDRASEKA 2013)	Modelos de agrupamento Fuzzy, MVS com kernels (linear, polinomial e RBF).	LIDC-IDRI	54
(AKRAM <i>et al.</i> , 2015)	Características de textura baseadas em técnicas estatísticas e MVS.	LIDC	40
(ESKANDARIAN BAGHERZADEF 2015)	Algoritmo genético e MVS.	LIDC	147
(SETIO <i>et al.</i> , 2016)	Muti-view redes convolucionais e MVS.	LIDC	888
(TERAMOTO <i>et al.</i> , 2016)	CNN, medidas de forma, classificador baseado em regras e MVS.	-	104

## 2.2 Diagnóstico

Nesta seção, serão apresentados brevemente trabalhos disponíveis na literatura, que tratam sobre o diagnóstico de nódulos pulmonares em maligno e benigno, sendo essa outra vertente explorada por nossos métodos.

A maioria dos trabalhos na área de diagnóstico de nódulos pulmonares mostra que para a classificação desse em maligno ou benigno são utilizados não só descritores de textura, mas também descritores que analisam a sua forma, devido à complexidade do problema. Nesse sentido, a seguir, tem-se um resumo de alguns desses trabalhos.

O objetivo da metodologia desenvolvida por Way *et al.* (2009) é realizar a classificação de nódulos pulmonares em maligno e benigno. As imagens usadas foram obtidas através de um método de contorno ativo. Um conjunto de imagens de 256 nódulos pulmonares (124 malignas e 132 benignas) extraídos de 152 pacientes foi utilizado para testes. Os descritores utilizados foram baseados em análises de forma e textura. Na classificação, utilizou-se o classificador LDA. O método alcançou uma acurácia total de 85%.

Em El-Baz *et al.* (2011) é proposto um método de diagnóstico de nódulos pulmonares em maligno e benigno baseado apenas na forma. A aquisição das imagens é feita através de uma delimitação da superfície usando análise esférica. A análise da forma 3D proposta pelos autores é realizada em cinco etapas: 1) segmentação do nódulo pulmonar com um limite deformável; 2) a triangulação de Delaunay é usada para construir um modelo 3D de malha da superfície; 3) em seguida, é feito um mapeamento desse modelo para uma esfera; 4) calculam-se as esferas harmônicas para a superfície e, por fim, 5) determinam-se o número de esferas harmônicas para delinear o nódulo. Os experimentos foram realizados usando um conjunto de 327 nódulos pulmonares (153 malignos e 174 benignos). A classificação alcançou uma acurácia de 93,6%.

O método de Kumar *et al.* (2011) propõe o desenvolvimento de um sistema CADx. Na primeira fase, a imagem de entrada é pré-processada e a região do nódulo é segmentada. A segunda etapa envolve o diagnóstico do nódulo com base na área e nos níveis de cinza da região do nódulo. Em seguida, é feita a redução de falsos positivos, enquanto mantém um elevado grau de verdadeiro positivo. O método proposto alcança uma precisão de 90%.

No trabalho de 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 usados para treinar uma rede neural com RBF. A partir dos resultados experimentais, foi encontrada uma precisão de 94,44%.

Em Orozco *et al.* (2012), é proposto um método para classificação de nódulos pulmonares. As regiões usadas para classificação foram selecionadas manualmente. Em seguida, calcularam-se dois espectros usando a *Discrete Cosine Transform* (2D-DCT) e *Fast Fourier Transform* (2D-FFT). Posteriormente, duas características de textura de natureza estatística foram extraídas do histograma calculado a partir dos espectros gerados. Por fim, foi usada a MVS para etapa de classificação. O método foi validado usando 75 amostras de nódulos. Após a classificação, os resultados alcançaram uma acurácia de 82,66%.

Em Nascimento *et al.* (2012), é proposta uma metodologia de caracterização de nódulos pulmonares, com o objetivo de se tornar uma ferramenta computacional utilizada para sugerir sobre a malignidade e a benignidade dos mesmos, atuando como uma segunda opinião junto ao especialista. 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 essa etapa, foi realizada a classificação pela MVS, onde foram obtidas taxas de sensibilidade de 85,64%, especificidade de 97,89% e acurácia de 92,78%.

Já Orozco *et al.* (2013) propõem uma alternativa computacional para classificar nódulos pulmonares por meio de TC. A novidade do método é a eliminação da fase de segmentação. Após a aquisição das imagens, oito características de textura foram extraídas a partir do histograma e da matriz de co-ocorrência de níveis de cinza (com quatro ângulos diferentes) para cada

imagem de TC. As características foram usadas para treinar um classificador MVS. Os resultados obtidos foram validados por um radiologista para finalmente se obter uma acurácia de 84%.

No método desenvolvido por Zhang *et al.* (2013) apresenta um método de classificação de nódulos pulmonares em *Low dose computed tomography* (LDCT). Em suma, o método proposto divide-se nas seguintes etapas: 1) rotulagem dos superpixel, isto é, rotula os pixels em objeto e fundo; 2) calcula-se uma curva de contexto, que converte o resultado da rotulagem em características do nódulo. Enquanto a primeira fase é responsável por extrair as principais estruturas de contextos anatômicos para cada tipo de nódulo, a curva de contexto fornece uma descrição para tipos diferentes de nódulos. Os resultados apresentados demonstram uma acurácia de 78%.

Em Krewer *et al.* (2013), recursos de textura e forma foram extraídos de nódulos pulmonares selecionados do conjunto de dados 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%.

O método proposto em Parveen e Kavitha (2014) consiste em quatro fases: 1) extração da região do pulmão; 2) segmentação; 3) extração de características e, finalmente, a classificação, em benigno e maligno. A identificação do pixel da lesão com o método de crescimento de região é utilizada para a segmentação das zonas focais dentro do pulmão. Para a extração de atributos, utiliza-se a matriz de co-ocorrência. As características extraídas são classificadas utilizando diferentes núcleos do MVS. Como resultados dos experimentos, foram obtidos uma sensibilidade de 91,38% e especificidade de 89,56%.

O trabalho de Dandil *et al.* (2014) foi desenvolvido usando TC para realizar a diferenciação entre tumores benignos e malignos. O sistema CADx projetado fornece segmentação de nódulos nos lóbulos usando *Self-Organizing Maps* e realiza a classificação entre nódulos benignos e malignos com o auxílio da Artificial Neural Network (ANN). Por fim, os resultados obtidos são de 90,63% para a acurácia, sensibilidade de 92,30% e 89,47% de especificidade.

No CADx desenvolvido por Orozco *et al.* (2015), tem como base as seguintes etapas: 1) a extração da região de interesse; 2) as transformadas

*Wavelet* são calculadas com um e dois níveis de decomposição; 3) seguindo, 19 características são extraídas. Após a extração, selecionam-se os descritores mais significativos, restando apenas 11. Por fim, é usado o classificador MVS. O trabalho foi validado em 45 TCs das bases ELCAP e LIDC. A acurácia total do método foi de 82% a sensibilidade de 90,9% e a especificidade de 73,91%.

Foi proposto por Kumar *et al.* (2015) um sistema CADx que usa características extraídas de um autoencoder para classificar nódulos pulmonares como maligno ou benigno. Foram usados 4.303 casos contendo 4.323 nódulos do conjunto de dados do *National Cancer Institute* (NCI) e LIDC para obter uma acurácia de 75,01% e sensibilidade de 83,35%.

Na metodologia CADx desenvolvida por Firmino *et al.* (2016), tem-se um método de diagnóstico de nódulo pulmonar baseada na probabilidade do mesmo ser maligno. Para isso, usou-se uma classificação baseado em regras e MVS. O banco de imagens usado para testes é o LIDC-IDRI. Os resultados da metodologia alcançaram uma sensibilidade de 93,1% com uma área sob a curva ROC de 0,91.

No CADx proposto em Dhara *et al.* (2016), os nódulos pulmonares são segmentados usando uma técnica semiautomática. Medidas de forma e textura são usadas para representar os nódulos pulmonares. O esquema de classificação proposto é validado no conjunto de nódulos com 891 nódulos pulmonares extraídos da base LIDC-IRI. O desempenho do classificador MVS é avaliada em termos da área sob a curva ROC. Os resultados alcançados pelo método proposto alcançaram uma curva ROC de 0,882.

Esses são exemplos de trabalhos que foram desenvolvidos para o diagnóstico de nódulo pulmonar. A Tabela 2.2 apresenta um resumo das abordagens. Alguns pontos, como: poucos casos de análise, discrepância entre os valores de sensibilidade e especificidade, são problemas comuns de sistemas dessa natureza. Sendo assim, os métodos propostos buscarão explorar essas deficiências com o intuito de melhorá-las, para isso, serão propostos métodos inovadores e com abordagens novas, com o intuito de analisar a forma e a textura dos nódulos pulmonares.

Tabela 2.2: Resumo dos trabalhos relacionados que descrevem métodos de diagnóstico de nódulo pulmonar.

Trabalho	Técnicas	Base	Amostra
(WAY <i>et al.</i> , 2009)	Características de forma usando <i>Gradient field</i> e características de raios, <i>stepwise, simplex optimization</i> , LDA e MVS.	Privado	256
(EL-BAZ <i>et al.</i> , 2011)	Características de forma usando <i>Spherical Harmonics</i> , mapeamento do modelo de esfera e classificador KNN	Privado	327
(KUMAR <i>et al.</i> , 2011)	Características de forma usando <i>bi-orthogonal wavelet</i> e classificador Fuzzy.	Privado	-
(ELIZABETH <i>et al.</i> , 2012)	Características de forma e textura e <i>radial basis function neural network</i>	Privado	-
(OROZCO <i>et al.</i> , 2012)	Atributos baseados em textura, seleção de medidas baseadas em correlação, KNN e MVS	NBIA-ELCAP	-
(NASCIMENTO <i>et al.</i> , 2012)	Descritores de textura usando índices de diversidade de <i>Shannon</i> e <i>Simpson</i> , <i>stepwise</i> e MVS	LIDC-IDRI	73
(OROZCO <i>et al.</i> , 2013)	Medidas baseadas no histograma e na matriz de co-ocorrência.	LIDC e NCI	4323
(ZHANG <i>et al.</i> , 2013)	Divisão de caminho, construção de contexto de curva, PCA, LDA, <i>scale-invariant feature transform</i> (SIFT) e MVS	ELCAP	50
(KREWER <i>et al.</i> , 2013)	Características de textura, seleção de atributos baseada em correlação e KNN	LIDC-IDRI	33
(PARVEEN; KAVITHA, 2014)	Características de textura baseadas na matriz de co-ocorrência e MVS	Privado	11
(DANDIL <i>et al.</i> , 2014)	Características de textura baseadas na matriz de co-ocorrência, PCA e ANN	Privado	128
(OROZCO <i>et al.</i> , 2015)	Transformada de <i>Wavelet</i> e medidas de textura baseadas na matriz de co-ocorrência.	ELCAP e LIDC	106
(KUMAR <i>et al.</i> , 2015)	Características baseadas na textura feita usando uma janela adaptativa.	ELCAP e NBIA	128
(FIRMINO <i>et al.</i> , 2016)	Descritores de textura e forma e classificação baseadas em regras e MVS	LIDC-IDRI	1109
(DHARA <i>et al.</i> , 2016)	Características baseadas na textura e forma e MVS.	LIDC-IDRI	891

# Fundamentação Teórica

---

Neste capítulo, será fornecida a base teórica que fundamenta esta tese. Será dado um enfoque geral, de modo a tornar compreensível a teoria que será utilizada na resolução dos problemas de redução de falsos positivos e diagnóstico dos VOIs de TC. As seções seguintes exploram as imagens de TC e suas características, propriedades do nódulo pulmonar, processamento de imagens digitais no âmbito da segmentação, extração de características de textura baseado em índices de diversidade filogenética, extração de características de forma baseada em descritores que exploram as propriedades morfológicas do VOI, algoritmos evolutivos e reconhecimento de padrões.

### 3.1 Fisiologia, Imagem e Patologia Pulmonar

O câncer de pulmão é um dos tumores malignos mais comuns do mundo. Segundo estimativas da Organização Mundial da Saúde (OMS), cerca de um milhão e setecentos mil casos novos são diagnosticados anualmente. O câncer pulmonar é caracterizado pela quebra dos mecanismos de defesa naturais do pulmão, a partir de estímulos carcinogênicos ao longo dos anos, levando ao crescimento desorganizado de células malignas. Esses fatores podem ocasionar o aparecimento de tumores, mostrados pela Figura 3.1. Além desses, outros achados podem estar contidos nos exames de TC.

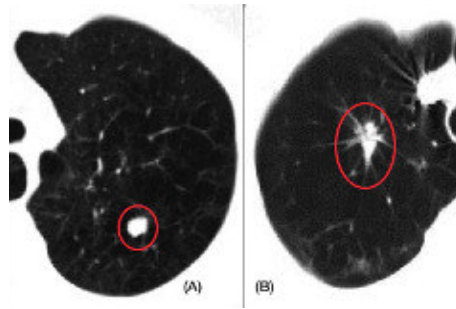


Figura 3.1: Exemplos de TC, onde é possível verificar o aparecimento de nódulos pulmonares: (A) TC com o nódulo benigno e (B) TC com nódulo maligno. Fonte: (ARMATO *et al.*, 2011).

### 3.1.1 Nódulo Pulmonar

Um nódulo é um conglomerado tecidual de até 3 cm, que se forma dentro ou na superfície do corpo, normalmente em resposta a lesões. A probabilidade de câncer varia muito com o tamanho dos nódulos nas diferentes populações. Estudos realizados nos Estados Unidos da América afirmam que cerca da metade dos nódulos com mais de 2 cm de diâmetro são malignos (DEWAN *et al.*, 1997). Já no cenário brasileiro, pela alta incidência de tuberculose, provavelmente esse percentual seja menor. De qualquer forma, a probabilidade de malignidade é diretamente proporcional ao tamanho do nódulo. Nos nódulos de diâmetro pequeno, a conduta do especialista atualmente é o acompanhamento durante um período de tempo para verificar o comportamento do nódulo. Cerca de 80% dos nódulos benignos possuem menos de 2 cm de diâmetro. No entanto, 15% dos nódulos malignos são menores que 1 cm e aproximadamente 42% são menores que 2 cm (ERASMUS *et al.*, 2000).

Tendões e músculos são ambos propensos ao desenvolvimento de nódulos em resposta às lesões. Nódulos podem também aparecer na pele ou nas camadas da pele, caso em que eles são geralmente notáveis, por serem visíveis. Órgãos, glândulas e outras partes do corpo podem também desenvolver nódulos (como mama, pulmão, próstata, entre outros) (HANSELL *et al.*, 2008).

Um nódulo benigno é basicamente um tumor que surge e não se espalha para outras partes do corpo. Os nódulos benignos tendem a crescer mais lentamente do que os malignos e são menos susceptíveis de causar problemas



de saúde. Mas isso não significa que o paciente poderá simplesmente esquecê-lo, deve haver um acompanhamento.

Câncer é o nome dado a um conjunto de mais de 100 doenças que têm em comum o crescimento desordenado de células que invadem os tecidos e órgãos, podendo causar a metástase<sup>1</sup> para outras regiões do corpo (OST *et al.*, 2003).

Como mostrado em Chen *et al.* (2011), a fase inicial do desenvolvimento do nódulo pulmonar é assintomática<sup>2</sup> e é formada por uma área pré-cancerosa não identificável por exames de imagem. Com o passar do tempo, ocorre o desenvolvimento do nódulo e, possivelmente, o espalhamento de suas células alteradas pela corrente sanguínea, desenvolvendo o processo de metástase.

### 3.1.1.1 Nódulo Pulmonar Solitário

Um Nódulo Pulmonar Solitário (NPS) é caracterizado por uma opacidade arredondada no pulmão com até 3 cm de diâmetro, envolta por parênquima pulmonar. Lesões pulmonares maiores que 3 cm de diâmetro são consideradas massas e muitas vezes malignas (FUJIMOTO; WISTUBA, 2014; GOULD *et al.*, 2001).

Dois padrões de margens de um nódulo são relativamente específicos para câncer. Um deles é o sinal corona radiata, que consiste em fios muito finos lineares que se estendem de 4 a 5 mm para fora do nódulo, originalmente descritos em tomógrafos simples, pois eles têm uma aparência espiculada em radiografias simples, como ilustrado na Figura 3.2, ao passo que uma borda mais lisa é mais sugestiva de um diagnóstico benigno (OST *et al.*, 2003).

Os padrões de calcificação são mais facilmente observados em TC do que em película simples de radiografias. Um padrão laminado ou central é típico de um granuloma<sup>3</sup>. Tem-se também os hamartomas, que são o tipo mais comum de tumor pulmonar benigno. Eles ocorrem principalmente em adultos, ainda que ocasionalmente surtem em crianças. Hamartomas possuem localização pulmonar periférica e padrões grosseiros de pipoca. Histologicamente, consiste geralmente de tecidos epiteliais e de outros tecidos, tais como gordura e cartilagem.

---

<sup>1</sup>É a disseminação do câncer para outros órgãos - quando as células cancerígenas desprendem do tumor primário e entram na corrente sanguínea ou no sistema linfático.

<sup>2</sup>Não apresenta sintomas.

<sup>3</sup>Lesão de aspecto proliferativo, caracterizado por um crescimento benigno do tecido.

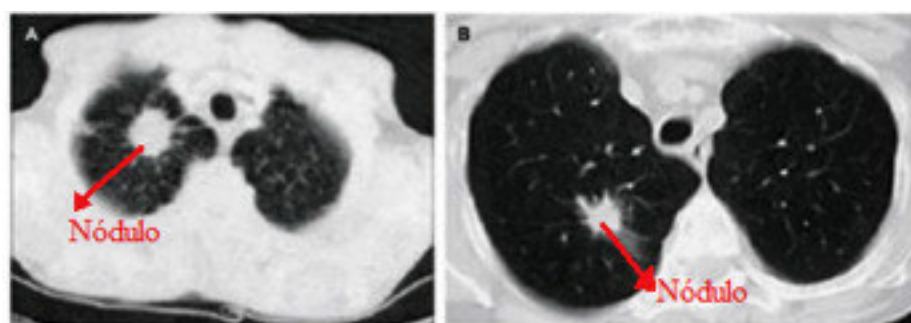


Figura 3.2: Aparência espiculada e não-espiculada de um nódulo pulmonar solitário.

A taxa de crescimento de um nódulo pode ser estimada se as imagens anteriores estiverem disponíveis para permitir uma medição precisa de alterações no seu tamanho (OST *et al.*, 2003). O tempo de duplicação do volume para os tumores malignos broncogênico é raramente inferior a um mês ou mais de um ano.

A utilização de estabilidade como um indicador de um processo benigno é baseada na medição precisa de crescimento e sobre a resolução da técnica de imagiologia utilizada (TAN *et al.*, 2003).

O nódulo é frequentemente descoberto por acaso, em uma radiografia ou tomografia de tórax feita por outras razões. Quando a lesão tem mais de 3 cm de diâmetro é chamada de massa; na presença de massa, a chance de câncer é alta (HANSELL *et al.*, 2008).

Nos trabalhos de Patil *et al.* (2014), Stewart e Wild (2014), percebe-se que 60% dos NPS são benignos. Dentre as causas benignas, destacam-se a tuberculose, as infecções por fungos, os hamartomas e os tumores benignos. As células são normais, mas são mal organizadas. No pulmão, o tipo mais frequente envolve um agrupamento nodular de cartilagem e gordura.

A preocupação da detecção de um nódulo pulmonar reside na possibilidade do mesmo ser um câncer (30%) ou um câncer proveniente de outro lugar que se disseminou para o pulmão (10%), de modo que as condições de cada ambiente influenciam nesses percentuais (PATIL *et al.*, 2014; STEWART; WILD, 2014).

De acordo com Chen *et al.* (2011), o tabagismo, o envelhecimento da sociedade, a urbanização, a poluição, o estilo de vida ruim e a industrialização são os principais fatores relacionados aos casos de câncer de pulmão. O tabagismo,

entretanto, representa cerca de 80% dos casos.

No entanto, o tema vem se revestindo de especial importância. Nas últimas décadas, principalmente com o advento da TC helicoidal e mais modernamente de múltiplos detectores (*multi-slice*), estes têm permitido cada vez mais o diagnóstico de nódulos antes invisíveis ao Raio X simples. Uma TC visualiza todas as características importantes do nódulo, o distingue melhor de alterações da parede torácica, pode identificar lesões adicionais, analisar melhor os ápices, regiões peri-hilares e ângulos costofrênicos, permitindo um melhor estudo do mediastino.

## 3.2 Tomografia Computadorizada

A Tomografia Computadorizada foi desenvolvida por Godfrey Hounsfield e Allan Cormack, em 1972 e o seu funcionamento consiste basicamente em uma máquina que emite raios-X na medida em que gira em volta do corpo do paciente, tal máquina denomina-se tomógrafo. Uma parte muito importante do tomógrafo é o *Gantry*. O *Gantry* é uma moldura circular, rotatória com um tubo de raios-X montado em um lado e um detector no lado oposto. Um feixe de raios-X em formato de leque gira o tubo de raios-X e o detector em torno do paciente. Conforme o aparelho roda, muitas imagens são tiradas em uma rotação resultando em uma imagem de seção transversal completa do corpo. Por fim, de acordo com a densidade dos tecidos, o feixe de raios-x será atenuado. Após essa interação, o feixe atinge os detectores que, por sua vez, são convertidos do estado analógico para o sistema computacional produzindo a imagem de TC (SARAVANAKUMAR *et al.*, 2014).

TC é um exame que permite a obtenção de imagens de cortes do corpo do paciente, sendo bastante utilizado como um exame médico de diagnóstico por imagem. Comparada à radiografia tradicional, a TC apresenta maior precisão e sensibilidade, além de não apresentar uma imagem com sobreposição de tecidos (FILHO *et al.*, 2013). Ela é excelente para detecção de alterações agudas ou crônicas no parênquima pulmonar, como as decorrentes de enfisema ou fibrose (doença pulmonar obstrutiva crônica, que causa a perda de capacidade respiratória e uma oxigenação insuficiente). É também usada na região do tórax

para diagnóstico de várias outras infecções.

Com a TC, é possível uma visualização tridimensional dos órgãos em que ela foi realizada, em especial no tórax. A TC proporciona uma maior flexibilidade nos estudos das imagens com maior número de detalhes e com maior número de níveis de cinza, se comparado com um exame de raios X. Essa seria então a grande vantagem do uso da TC.

O resultado visual da TC é monocromático, ou seja, são mostrados apenas os vários níveis de cinza, indo do totalmente preto ao branco e, mesmo assim são muito mais numerosas que as variações de níveis de cinza dos raios X convencionais. Todas as estruturas presentes na TC puderam ser proporcionalmente quantificadas em unidades de densidade relativas ao padrão, chamadas de Unidades Hounsfield (UH), em outras palavras, a UH representa os níveis de cinza de uma imagem de TC.

### 3.3 Arquivo DICOM

Com a introdução da TC, seguido por outras modalidades digitais de diagnóstico por imagem na década de 1970 e o crescente uso de computadores para aplicações clínicas, o *National College of Radiography* (NCR) e a *National Electrical Manufacturers Association* (NEMA) reconheceu a emergente necessidade de um método padrão para a transferência de imagens e informações associadas entre dispositivos fabricados por diversos fornecedores (VARMA, 2012).

O ACR e a NEMA formaram uma comissão conjunta em 1983 para desenvolver um padrão para:

- Promover a comunicação de informação de imagem digital, independentemente do fabricante do dispositivo;
- Facilitar o desenvolvimento e a expansão de arquivamento de imagens e sistemas de comunicação; e
- Permitir a criação de bases de informações de dados de diagnóstico que podem ser acessados por uma grande variedade de dispositivos distribuídos geograficamente.

### 3.3.1 Padrão DICOM

Este padrão é atualmente denominado Digital Imaging Communications in Medicine (DICOM), que incorpora uma série de melhorias importantes para versões anteriores do padrão ACR-NEMA (VARMA, 2012).

- É aplicável a um ambiente de rede;
- Aplicável a um ambiente de comunicação *off-line*;
- Especifica como os dispositivos em conformidade com a norma reagem a comandos e dados a serem trocados;
- Ele especifica os níveis de conformidade. DICOM explicitamente descreve como um programador deve estruturar uma declaração de conformidade para selecionar opções específicas;
- Ele é estruturado como um documento de peça múltipla;
- Introduz objetos de informação explícita não só para imagens e gráficos, mas também para formas de onda, relatórios, impressão, etc.; e,
- Ele especifica uma técnica estabelecida para identificar exclusivamente qualquer objeto de informação. Isso facilita as definições inequívocas de relações entre objetos de informação à medida que são postas em prática em toda a rede.

Cada arquivo DICOM, ao contrário da maioria dos outros padrões para imagens médicas, é composto por um cabeçalho e um corpo. O cabeçalho contém informações úteis na interpretação do conteúdo do corpo do arquivo. Nele, pode-se encontrar o tamanho da imagem, o nível de compressão, a data do exame, etc. O corpo carrega a imagem propriamente dita e pode ser armazenado tanto em formato não processado quanto compactado em formato JPEG (SOUSA *et al.*, 2010).

## 3.4 Base LIDC-IDRI

Nesta tese, foram usadas as imagens disponibilizadas pelo *National Cancer Institute of EUA* (NCI), criadas a partir de um repositório de imagens de câncer,

resultado da formação de um consórcio de instituições conhecido como *Lung Image Database Consortium* (LIDC).

A LIDC é um grupo que procura estabelecer um formato padrão e processos de gerenciamento de imagens de pulmão, relatórios técnicos e dados clínicos necessários para o desenvolvimento, treinamento e evolução de algoritmos para detectar e diagnosticar câncer de pulmão.

A base utilizada neste trabalho é a LIDC-IDRI (ARMATO *et al.*, 2011), disponibilizada recentemente como resultado de uma associação entre o *Lung Image Database Consortium* e a *Image Database Resource Initiative*. Essa parceria promoveu um aumento substancial da base LIDC, resultando em um total de 1.018 casos. Os exames disponibilizados por essa base apresentam as mesmas características da LIDC original.

Todas as imagens estão no formato DICOM e possuem até 16 bits por *voxel*<sup>4</sup>. A base conta com um arquivo que contém informações sobre marcações de nódulos realizadas por quatro especialistas. Esse arquivo também possui outras informações dos nódulos, como: taxa de malignidade, dificuldade de detecção, textura, nível de calcificação, esfericidade, entre outras. As imagens possuem dimensões de 512 x 512 (altura e largura) com quantidades variadas de imagens para cada exame. As TCs foram adquiridas em tomógrafos distintos. Esse fator dificulta a detecção, pois o contraste pode ser variado.

Essa base dispõe de um arquivo individual para cada exame que contém informações sobre cada nódulo presente no exame. Nessa base, encontram-se diversos tipos de nódulos, dentre eles tem-se: pequenos (diâmetro maior que 3 e menor que 10 mm), conectados a vasos sanguíneos, justapleurais e nódulos grandes (diâmetro maior que 10 e menor que 30 mm). Existe também a presença de massas, que são estruturas com diâmetro maior que 30 mm.

Nesse arquivo, são encontradas as informações mais valiosas para concretização de qualquer método de detecção ou diagnóstico de nódulo pulmonar. Nele, encontram-se as marcações realizadas previamente por quatro especialistas. Essas marcações indicam o contorno do nódulo pulmonar e as fatias de TC que o compõe. Assim é possível afirmar ao certo quais candidatos a nódulos são de fato nódulos pulmonares.

---

<sup>4</sup>Representação tridimensional de um pixel.

Além dessas marcações, os arquivos dispõem de informações referentes aos nódulos pulmonares, que são:

- **Sutileza** (*Subtlety*) - mede a dificuldade de detecção que os nódulos possuem, medidos em uma escala de 1 a 5, isto é, quanto menor o valor de sutileza, maior o grau de dificuldade de detecção;
- **Estrutura interna** (*Internal structure*) - composição interna do nódulo dividido em: tecido mole, líquido, gordura, ar;
- **Calcificação** (*Calcification*) - padrão ou nível de calcificação caso tenha, também medidos em uma escala de 1 a 5;
- **Esfericidade** (*Sphericity*) - a forma tridimensional do nódulo em termos da sua quão esférico é o nódulo, distribuídos em uma escala de 1 a 5, sendo que, apenas a escala 1, 3 e 5 estão bem definidas pelos especialistas;
- **Espiculação** (*Spiculation*) - grau de espiculação do nódulo, em uma escala de 1 a 5;
- **Textura** (*Texture*) - textura interna do nódulo, marcada em uma escala de 1 a 5;
- **Malignidade** (*Malignancy*) - avaliação do especialista sobre a probabilidade de malignidade do nódulo, medida em uma escala de 1 a 5.

Essas marcações acerca do contorno do nódulo e de suas características são encontradas apenas para nódulos que possuem o diâmetro maior que 3 mm e menor que 30 mm. Para nódulos menores, é encontrada apenas a marcação referente ao centro de massa do mesmo.

### 3.5 Sistemas Computacionais de Auxílio à Detecção e ao Diagnóstico

A avaliação do radiologista é baseada em subjetividade, estando sujeita a variações intra e inter pessoais, bem como perda de informação devido a natureza

sutil do VOI, baixa qualidade da imagem, sobreposição de estruturas, fadiga visual ou distração. Além desses, o crescente avanço do processamento de imagem digital (hardware ou software), o reconhecimento de padrões e o uso de inteligência artificial proporcionam um cenário favorável para o desenvolvimento de mecanismos computacionais que, no uso destas tecnologias, proporcionem aos especialistas uma forma de auxílio.

Assim sendo, os sistemas de Auxílio à Detecção e Diagnóstico tornam-se ferramentas de extrema utilidade no auxílio ao especialista. Já foi demonstrado, por exemplo, que uma dupla leitura (por dois radiologistas) pode aumentar a sensibilidade do diagnóstico (AZEVEDO; CONCI, 2003). Desse modo, os sistemas de auxílio podem atuar como um segundo especialista ou uma segunda opinião para a tomada de decisão, com a finalidade de melhorar a consistência da interpretação da imagem radiológica.

Basicamente, existem dois tipos de aplicações desses sistemas: um é o auxílio à detecção de lesões (CAD) e o outro é o auxílio ao diagnóstico de lesões (CADx). De forma geral, o CAD divide-se em quatro etapas, sendo elas: aquisição de imagens, melhoramento, segmentação e, por fim, redução de falsos positivos (FILHO *et al.*, 2013). Já o CADx tem sua função voltada para o diagnóstico, isto é, diferenciar entre padrões de malignidade e benignidade.

Na literatura, existem CADs para todos os tipos de exames e várias partes do corpo, como crânio, tórax, abdômen, osso, sistema vascular, entre outros. Porém, os principais objetos de pesquisa para o desenvolvimento de sistemas CAD e CADx têm sido nos âmbitos da tomografia computadorizada, mamografia e angiografia.

O CAD e o CADx se complementam, uma vez que o procedimento comum do especialista é detectar a região suspeita, validar - através de algum meio - se essa região está clinicamente comprometida e, por fim, se essa região apresenta comportamento maligno ou benigno. Independentemente do emprego, é consenso na literatura o alerta para a não utilização de sistemas CAD como único meio de detecção e diagnóstico. Isto decorre porque o seu objetivo principal é contribuir para superar limitações humanas na realização de tarefas repetitivas, tais como distrações e fadigas, comuns na análise de grande quantidade de exames.



## 3.6 Processamento Digital de Imagens

De uma forma sucinta, pode-se aferir que o processamento digital de imagens é definido como sendo o conjunto de técnicas computacionais que manipulam uma imagem de entrada em uma saída, sendo que, na maioria dos casos, a saída é outra imagem digital. Dessa maneira, é possível, além de tratar aspectos visuais e estruturais, fornecer outros elementos que possibilitem a interpretação visual e computacional da imagem.

Uma imagem em duas dimensões (2D) pode ser definida como uma função bidimensional,  $f(x,y)$ , em que  $x$  e  $y$  são coordenadas espaciais e a amplitude de  $f$  em qualquer coordenada  $(x,y)$  é chamada de intensidade ou nível de cinza da imagem, para esse ponto. Assim, quando  $x$ ,  $y$  e  $f$  são quantidades finitas e discretas, chamamos de imagem digital (GONZALEZ; WOODS, 2006).

Os avanços tecnológicos ocorridos nas últimas décadas, sejam eles de hardware ou software, estão possibilitando um leque de aplicações cada vez maior na área de processamento digital de imagens. Em decorrência disso, muitas pesquisas são desenvolvidas com o intuito de aprimorar e expandir mais ainda ferramentas que possuem, em sua essência, o processamento de imagens. Além disso, conceitos comuns no processamento de imagens estão sendo amplamente utilizados em outras áreas de conhecimento, tais como: astronomia, física, arqueologia, biologia, defesa militar, aplicações industriais, etc. A medicina é uma das diversas áreas que tem sido beneficiada com um grande número de pesquisas relacionadas ao processamento de imagens médicas, sobretudo no desenvolvimento de ferramentas CAD/CADx.

Segundo Gonzalez e Woods (2006), um sistema clássico de processamento digital de imagens pode ser agrupado em cinco etapas: aquisição, pré-processamento, segmentação, extração de características e classificação.

Na primeira etapa, tem-se a aquisição de imagens. É nessa etapa que as imagens são capturadas e representadas de forma computacional para serem interpretadas pela etapa seguinte.

No pré-processamento, são realizados procedimentos capazes de proporcionar um melhoramento nos aspectos visuais e estruturais da imagem. Dessa maneira, consegue-se aumentar contraste, eliminar ruídos, etc.

A etapa de segmentação consiste em dividir a imagem em objeto(s) e

fundo. Em outras palavras, essa etapa consiste em técnicas que de alguma maneira consigam formar padrões de agrupamento, gerando sub-regiões que possuem entre si alguma similaridade.

A representação e descrição, conhecida também por extração de características, tem como objetivo representar, através de valores, uma imagem ou partes dela. Estes valores, são características fundamentais que representam propriedades contidas nas imagens.

E por fim, tem-se a etapa de reconhecimento de padrões. Nessa etapa, os valores obtidos na etapa de extração de características são os insumos para que uma técnica de aprendizado de máquina possa então discernir entre possíveis padrões contidos em um grupo de imagens.

Para o desenvolvimento desta tese, foram utilizadas as etapas de segmentação, extração de características e reconhecimento de padrões, com base no modelo apresentado por Gonzalez e Woods (2006).

### 3.6.1 Segmentação

No contexto da segmentação, usou-se o algoritmo de Otsu (YANG *et al.*, 2012). Esse algoritmo é uma técnica destinada a encontrar um limiar ideal com base no histograma da imagem. Dessa forma, a imagem é dividida em duas partes, de tal maneira que a semelhança entre as duas partes é a menor possível.

A partir de uma imagem de nível cinza com pixels  $N$  e  $L$ , a probabilidade de ocorrência de níveis de cinza  $i$  na imagem é mostrado na Equação 3.1.

$$p_i = \frac{f_i}{N} \quad (3.1)$$

onde  $f_i$  é a frequência de ocorrência do nível de cinza  $i$ , sendo  $i = 1, 2, \dots, L$ .

Os pixels são divididos em duas classes,  $C_1$  e  $C_2$ , com níveis de cinza  $(1, 2, \dots, T)$  e  $(T + 1, T + 2, \dots, L)$ , respectivamente. As distribuições de probabilidade das classes são representadas nas Equações 3.2 e 3.3,

$$C_1 = \frac{p_1}{\omega_1(t)}, \dots, \frac{p_t}{\omega_1(t)} \quad (3.2)$$

$$C_2 = \frac{p_{t+1}}{\omega_2(t)}, \frac{p_{t+2}}{\omega_2(t)}, \dots, \frac{p_L}{\omega_2(t)} \quad (3.3)$$

em que as intensidades de cada uma das classes  $\omega_1(t)$  e  $\omega_2(t)$  são definidas nas Equações 3.4 e 3.5:

$$\omega_1(t) = \sum_{i=1}^t p_i \quad (3.4)$$

$$\omega_2(t) = \sum_{i=t+1}^L p_i \quad (3.5)$$

As médias para cada uma das classes  $C_1$  e  $C_2$ , representadas por  $\mu_1(t)$  e  $\mu_2(t)$ , respectivamente, são definidas pelas Equações 3.6 e 3.7:

$$\mu_1(t) = \sum_{i=1}^t \frac{i * p_i}{\omega_1(t)} \quad (3.6)$$

$$\mu_2(t) = \sum_{i=t+1}^L \frac{i * p_i}{\omega_2(t)} \quad (3.7)$$

A média total da intensidade da imagem  $\mu_T$  é definida nas Equações 3.8 e 3.9.

$$\mu(T) = \omega_1 * \mu_1 + \omega_2 * \mu_2 \quad (3.8)$$

$$\omega_1 + \omega_2 = 1 \quad (3.9)$$

Após os cálculos das médias de cada classe e da média total da imagem, o algoritmo de Otsu define a maior variância entre as classes através da Equação 3.10, isto é,  $1 \leq t \leq L$ . E por fim, na Equação 3.11, é definido o valor de  $t^*$ , que representa o limiar ideal.

$$\delta_B^2 = \omega_1 * (\mu_1 - \mu_T)^2 + \omega_2 * (\mu_2 - \mu_T)^2 \quad (3.10)$$

$$t^* = Max(\delta_B^2(t)) \quad (3.11)$$

O algoritmo de Otsu foi utilizado nesta tese para delimitação de regiões internas aos VOIs.

### 3.6.2 Extração de Características

Uma das tarefas mais complexas na análise de imagens está na definição de um conjunto de características que possam descrever de maneira concreta cada região contida em uma imagem, de modo a serem utilizados em processos de mais alto nível (PEDRINI; SCHWARTZ, 2008). Em outras palavras, a etapa de caracterização e representação da imagem consiste em uma etapa de fundamental importância no modelo clássico do processamento de imagem digital proposto por (GONZALEZ; WOODS, 2006). Através de modelos/métodos matemáticos, as propriedades da imagem são representadas em formato de atributos, para então, servirem de entrada para uma posterior análise de reconhecimento e classificação de padrões.

Em linhas gerais, a etapa de extração de características pode ser dividida em duas categorias de análises, sendo elas por: textura e forma. Em uma análise por textura, o objetivo geral é descrever aspectos da imagem no que diz respeito a suavidade, rugosidade e regularidade (GONZALEZ; WOODS, 2006). Já em uma análise baseada na forma da imagem, o intuito é extrair informações que mensuram sobre propriedades morfológicas da imagem.

Com base no exposto, nesta tese, são apresentadas métodos que buscam explorar estas vertentes da extração de características. Para tal, as seções que se seguem apresentam as fundamentações dos descritores propostos para cada tipo de análise, textura e forma.

## 3.7 Análise por Textura

A definição de textura encontrada na literatura é descrita de diversas formas. Segundo Haralick *et al.* (1973), textura é definida como a característica de uma região relacionada a coeficientes de uniformidade, densidade, aspereza, regularidade, intensidade, dentre outras características da imagem. Para Gonzalez e Woods (2006), a textura é intuitivamente descrita por medidas que quantificam suas propriedades de suavidade, rugosidade e regularidade. Baseado num conceito bidimensional, a textura é caracterizada pela dimensão que contém as propriedades primitivas da tonalidade e a outra corresponde aos relacionamentos espaciais entre elas.

A análise de textura é relevante em imagens digitais, uma vez que possibilita distinguir regiões da imagem que apresentam as mesmas características de padrões (AZEVEDO; CONCI, 2003). As três abordagens principais usadas em processamento de imagens para a análise de texturas são: estrutural, espectral e estatística (GONZALEZ; WOODS, 2006).

A abordagem estrutural considera que texturas são compostas de primitivas dispostas de forma aproximadamente regular e repetitiva, conforme regras bem definidas. A abordagem espectral é baseada em propriedades do espectro de Fourier, sendo utilizada principalmente na detecção de periodicidade global em uma imagem através da identificação de picos de alta energia no espectro. A abordagem estatística define a textura como um conjunto de medidas locais extraídas do padrão, favorecendo a descrição de imagens através de regras estatísticas que regem tanto a distribuição quanto a relação entre os diferentes níveis de cinza.

Nesta tese, foram utilizados para caracterizar a textura dos VOIs de TCs os índices de diversidade filogenética, que são medidas estatísticas.

### 3.7.1 Índices de Diversidade e Diversidade Filogenética

A diversidade é um termo frequentemente usado em Ecologia. O objetivo de um índice de diversidade é descrever a variedade de espécies presentes em uma comunidade ou região (MAGURRAN, 2004). Como mostrado por Magurran (2004), o conceito de comunidade é apresentado como um conjunto de espécies que ocorrem em um determinado lugar e tempo. 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.

A diversidade é medida através de duas variantes: a *riqueza de espécies*, que representa o número de espécies contidas em determinada região e a *abundância relativa*, que refere-se ao número de indivíduos de uma determinada espécie existente em uma região (PIANKA, 2011).

Para se usar índices de diversidade sobre imagens, faz-se necessário realizar uma analogia entre suas propriedades, isto é, a comunidade é representada por uma imagem ou região da mesma, as espécies são os níveis de cinza e os

indivíduos sendo os *pixels* ou *voxels* de uma determinada espécie. A Figura 3.3 mostra um exemplo dessa analogia.

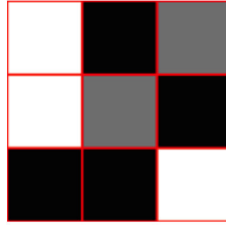


Figura 3.3: Exemplo de uma imagem que possui 3 níveis de cinza (espécies) que é o preto (P), cinza (C) e branco (B). A quantidade de pixels (indivíduos) de B é 3, C é 2 e P é 4.

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 (BAXEVANIS; OUELLETTE, 2004). 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 3.4, tem-se um exemplo de árvore filogenética, onde se verifica o relacionamento genético entre espécies de macacos e a espécie humana, em que é possível observar que o homem e o chimpanzé são geneticamente mais próximos que os outros pares presentes na árvore (BAXEVANIS; OUELLETTE, 2004).

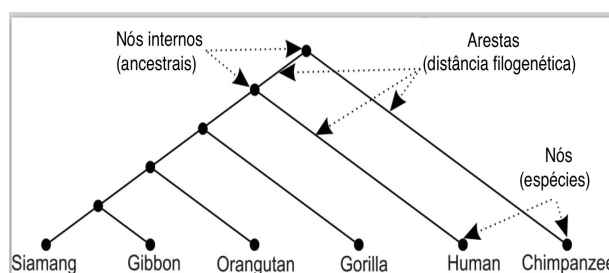


Figura 3.4: Exemplo de árvore filogenética para alguns primatas adaptada de (BAXEVANIS; OUELLETTE, 2004)

De maneira geral, a diversidade não pode ser medida apenas com a utilização de dados, como a abundância e a riqueza de espécies; e cada vez

mais o parâmetro filogenético vem sendo inserido nesse cálculo (PIENKOWSKI *et al.*, 1998). A diversidade filogenética é uma medida da diversidade de uma comunidade que incorpora as relações filogenéticas das espécies (MAGURRAN, 2004). A combinação de abundância das espécies com a proximidade filogenética para gerar um índice de diversidade é denotada diversidade taxonômica (SILVA; BATALHA, 2006). Segundo Vandamme *et al.* (1996), a taxonomia é a ciência que lida com a classificação (criação de novos táxons<sup>5</sup>), identificação (alocação de linhagens dentro de espécies conhecidas) e nomenclatura.

No trabalho de Pienkowski *et al.* (1998), foi desenvolvido um método para averiguar a diversidade taxonômica sensível a alterações ambientais e adequado para avaliar as diferenças entre comunidades. Uma comunidade em que as espécies estão distribuídas em vários gêneros deve apresentar uma diversidade maior que uma comunidade em que a maioria das espécies pertence a um mesmo gênero (MAGURRAN, 2004).

A diversidade taxonômica baseia-se no conjunto das distâncias entre pares de espécies acumuladas a partir das árvores taxonômicas. A Figura 3.5 ilustra uma árvore taxonômica em que as folhas são as espécies e a soma da quantidade de arestas que ligam determinado par de espécies, representado na matriz, indica a distância filogenética existente.

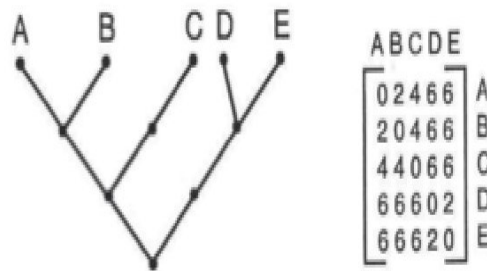


Figura 3.5: Exemplo de uma árvore taxonômica com sua matriz de distância entre as espécies.

Uma das formas de representar a árvore filogenética é através do dendrograma, que é um diagrama representativo das relações ancestrais entre organismos. Para esta tese, foi utilizada essa forma de representação da imagem.

<sup>5</sup>Unidade taxonômica, essencialmente associada a um sistema de classificação científica.

Esse tipo de árvore é mostrado na Figura 3.6, que descreve a sequência evolutiva de alguns primatas.

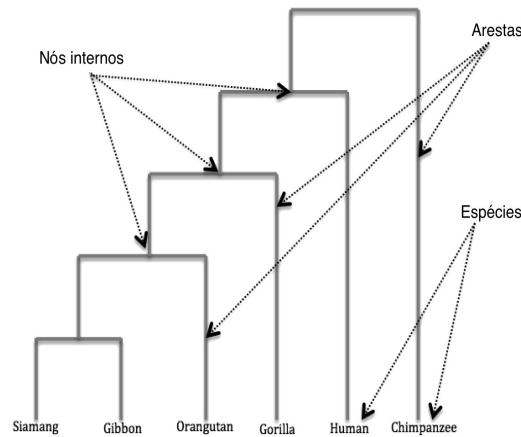


Figura 3.6: Árvore filogenética na forma de dendrograma.

Árvores filogenéticas combinadas com os índices de diversidade filogenética são usados na biologia para comparar padrões de comportamento entre as espécies de diferentes áreas. Para implementar essa ideia, o primeiro passo é fazer uma correspondência entre os termos usados na biologia e os utilizados neste trabalho. A Tabela 3.1 mostra essa correspondência.

Tabela 3.1: Correspondência entre os termos da biologia e os métodos propostos.

Biologia	Tese
Comunidade	VOIs de TC
Espécies	Número de UH encontrados no VOI
Riqueza de espécie: número de espécies encontradas em uma certa região	Número de <i>voxels</i> presentes no VOI
Indivíduos	Quantidade de <i>voxels</i> de uma determinada espécie contida em um VOI
Abundância relativa: número de indivíduos de uma determinada espécie existentes numa determinada área	Número de <i>voxels</i> encontrados no VOI que possuem o mesmo valor UH

Os índices de diversidade filogenética foram escolhidos devido ao seu



potencial na caracterização de uma determinada região/imagem. A riqueza de detalhes obtida com cada grupo de índice é fundamental para composição dos descritores propostos nesta tese. É importante mencionar que estes índices se complementam, isto é, um grupo de índice é capaz de mensurar alguma propriedade que outro grupo não consiga, como por exemplo, os índices de diversidade filogenética baseados na riqueza de espécies e o grupo de índices baseados na distância entre pares de espécies. O primeiro, é capaz de mensurar propriedades diretamente relacionadas a espécie, como a sua abundância relativa, isto é, quantidade de indivíduos de uma espécie. Já o segundo, é capaz de mensurar a cerca das relações de parentesco que determinadas espécies possuem, como por exemplo, a quantidade de ancestrais comuns que existem entre determinadas espécies.

### 3.7.1.1 Índices de Diversidade Filegenética Baseados na Riqueza de Espécies

A relação entre dois organismos escolhidos aleatoriamente em uma filogenia existente em uma comunidade é apresentada pelos índices de diversidade taxonômica ( $\Delta$ ) e distinção taxonômica ( $\Delta^*$ ) (PIENKOWSKI *et al.*, 1998). Esses índices possuem três fatores essenciais para sua aplicação: número de espécies, número de indivíduos, bem como a estrutura de ligação das espécies (número de arestas).

O índice de diversidade taxonômica ( $\Delta$ ) considera a abundância das espécies e a relação taxonômica entre eles. Portanto, o seu valor expressa a distância taxonômica média entre quaisquer dois indivíduos, escolhidos aleatoriamente em uma amostra (PIENKOWSKI *et al.*, 1998). Esse índice é definido pela Equação 3.12.

$$\Delta = \frac{\sum \sum_{i < j} w_{ij} x_i x_j}{[n(n-1)/2]} \quad (3.12)$$

onde  $x_i$  ( $i = 0, \dots, s$ ) é a abundância (número de *voxels*) da  $i$ -ésima espécie,  $x_j$  ( $j = 0, \dots, s$ ) representa a abundância de  $j$ -ésima,  $s$  indica o número de espécies,  $n$  é o total do número de indivíduos e  $w_{ij}$  é a distância entre a espécie  $i$  para a espécie  $j$  na classificação taxonômica.

O índice de distinção taxonômico ( $\Delta^*$ ), por sua vez, representa a distância taxonômica média entre dois indivíduos, com a restrição de que pertençam a espécies diferentes (PIENKOWSKI *et al.*, 1998). A Equação 3.13 define esse índice.

$$\Delta^* = \frac{\sum \sum_{i < j} w_{ij} x_i x_j}{\sum \sum_{i < j} x_i x_j} \quad (3.13)$$

onde  $x_i$  ( $i = 0, \dots, s$ ) é a abundância (número de *voxels*) da  $i$ -ésima espécie,  $x_j$  ( $j = 0, \dots, s$ ) é a abundância da  $j$ -ésima espécie,  $s$  indica o número de espécies e  $w_{ij}$  é a distância entre as espécies  $i$  e  $j$ .

### 3.7.1.2 Índices de Diversidade Filogenética Baseados na Distância entre Pares de Espécies

Muitos estudos ecológicos, especialmente os de grande escala, dependem da riqueza de espécies como medida de biodiversidade. No entanto, o uso de riqueza de espécies como o único reflexo da biodiversidade pode ser de valor limitado, uma vez que trata todas as espécies como sendo iguais e não leva em consideração as relações filogenéticas (VANE-WRIGHT *et al.*, 1991).

Em Sechrest *et al.* (2002) foi mostrado que pontos de biodiversidade em todo o mundo mostram maiores quantidades de história evolutiva do que se conseguiria baseado apenas na riqueza das espécies. O trabalho de Heard e Mooers (2000) demonstra que as relações filogenéticas são um dos fatores mais importantes que determinam a extinção de espécies. E em Strauss *et al.* (2006) foi apresentado que o grau de distância filogenética pode determinar o sucesso de invasão de espécies e subespécies exóticas. Por isso, a informação filogenética pode ser um melhor indicador do valor de preservação do que unicamente a riqueza de espécies. A aplicação das informações obtidas por relações filogenéticas representam uma abordagem promissora (WEBB *et al.*, 2002).

Os estudos que verificam as relações de distância entre os pares são baseadas em uma matriz de distância entre todas as espécies de uma comunidade. As distâncias podem basear-se em diferenças morfológicas ou funcionais (IZSAKI; PAPP, 1995), no comprimento dos ramos das relações filogenéticas baseadas em dados moleculares (SOLOW *et al.*, 1993; PAVOINE *et al.*, 2005) ou, se os comprimentos dos ramos não são conhecidos, sobre o número de nós que separam

cada par de espécies (FAITH, 1992). Os valores dentro da matriz de distância podem ser interpretados como a distinção entre cada par de espécies ou de cada espécie em particular para todas as outras (RAO, 1982; IZSAK; PAPP, 2000).

Esses índices baseiam-se nas distâncias entre pares de espécies. A entropia quadrática intensiva (IZSAK; PAPP, 2000); a entropia quadrática extensiva (IZSAK; PAPP, 2000); a distinção taxonômica média (CLARKE; WARWICK, 1998); a distinção taxonômica total (CLARKE *et al.*, 2001); e a diversidade pura (WEITZMAN, 1992; FAITH, 1994).

A entropia quadrática intensiva ( $J$ ) foi proposta inicialmente por Izsak e Papp (2000) com intuito de realizar uma possível ligação entre os índices de diversidade e os índices de medidas de biodiversidade. Quando se tem os mesmos valores de abundância (hipotética ou formal), o índice  $J$  (Equação 3.14) será uma função que representa o número de espécies e as suas relações taxonômicas. Dessa forma, expressando a distância taxonômica média entre duas espécies escolhidas aleatoriamente, as relações entre as espécies influenciam no valor de  $J$ , ao contrário de outros índices de diversidade (IZSAK; PAPP, 2000).

$$J = \left[ \sum d_{i,j} \right] / s^2 \quad (3.14)$$

onde  $d_{i,j}$  representa a distância entre as espécies  $i$  e  $j$  e  $s$  representa o número de espécies.

Em relação às propriedades do índice, *monotocidade* é uma propriedade geralmente necessária para as medidas de diversidade. Dessa forma, representando uma medida por  $I$ , essa propriedade significa que:  $I(A \cup \{x\}) > I(A)$ . Isso garante que o valor do índice irá aumentar, adicionando uma nova espécie  $x$  a um conjunto de espécies  $A$ . O índice  $J$  não irá satisfazer esse requisito. Ou seja, o índice  $J$  não é um índice de diversidade ideal. Para resolver isso, foi proposto, a partir do índice  $J$ , a entropia quadrática extensiva  $F$  (IZSAK; PAPP, 2000), que representa a soma das diferenças das espécies. A *monotocidade* é aplicada ao índice  $F$ , pois, para qualquer conjunto de espécies  $A$  e uma nova espécie  $x$ , o valor será alterado. A Equação 3.15 define o índice  $F$ .

$$F = \sum d_{i,j} \quad (3.15)$$

onde  $d_{i,j}$  representa a distância entre as espécies  $i$  e  $j$ .

Os índices de distinção taxonômica média (*AvTD*) (CLARKE; WARWICK, 1998) e distinção taxonômica total (*TTD*) (CLARKE *et al.*, 2001) foram originalmente desenvolvidos com base em relações taxonômicas. No entanto, eles podem ser facilmente adaptados à informação filogenética (SCHWEIGER *et al.*, 2008). *AvTD* indica a distância taxonômica média entre quaisquer duas espécies escolhidas aleatoriamente (CLARKE; WARWICK, 1998). Já o *TTD* representa a distinção média filogenética somada ao longo de todas as espécies. As Equações 3.16 e 3.17 definem esses índices, respectivamente.

$$AvTD = \left[ \sum_{i < j} \sum d_{ij} \right] / [s(s-1)/2] \quad (3.16)$$

$$TTD = \sum_i \left[ \left( \sum_{j \neq i} d_{ij} \right) / (s-1) \right] \quad (3.17)$$

Sendo que, para as Equações 3.16 e 3.17,  $d_{i,j}$  representa a distância entre as espécies  $i$  e  $j$  e  $s$  representa o número de espécies.

Por fim, tem-se o índice de diversidade pura ( $D_D$ ) (Equação 3.18), que é responsável por averiguar o valor da distância de uma espécie para seu vizinho mais próximo (FAITH, 1994; WEITZMAN, 1992).

$$D_D = \sum d_{i \min} \quad (3.18)$$

onde  $d_{i \min}$  indica a distância do vizinho mais próximo da espécie  $i$  para todas as outras espécies.

### 3.7.1.3 Índices de Diversidade Filogenética Baseados na Topologia

O estudo realizado por Vane-Wright *et al.* (1991) foi um dos primeiros a propor a aplicação de métodos baseados na topologia, que reflete a ordem de ramificação filogenética dentro de um grupo. Nessa abordagem, cada espécie de uma comunidade é ponderada pelo número de nós entre a espécie e a raiz da árvore filogenética (dendrograma). Dessa maneira, as espécies com os maiores pesos são as que possuem maior distância da raiz.

Os índices, soma básica dos pesos ( $Q$ ) e soma dos pesos normalizados ( $W$ ), foram usados para descrever a textura dos VOIs. Estes se baseiam na topologia da árvore que relaciona as espécies presentes em toda a comunidade, com isso esses índices mencionam o grau de parentesco (similaridade) entre as espécies (KEITH *et al.*, 2005; POSADAS *et al.*, 2001; VANE-WRIGHT *et al.*, 1991).

O índice  $Q$  representa o quociente entre a totalidade de nós para todo o grupo dividido pelo número de nós entre a raiz e uma espécie, ou seja, a soma das contribuições de cada espécie para a diversidade. O índice  $Q$  é definido na Equação 2.19. O índice  $W$ , definido pela Equação 2.20, indica o peso normalizado para cada espécie, isto é, os valores de  $Q$  de cada espécie dividido pelo valor de  $Q_{min}$ .

$$\begin{aligned} Q &= \sum Q_i \\ Q_i &= I/I_i \\ I &= \sum I_i \end{aligned} \tag{3.19}$$

onde  $Q_i$  representa o quociente do total de nós dos caminhos partindo da raiz até todas as espécies da árvore filogenética por  $I_i$ , que é o número de nós entre a raiz e a espécie  $i$ .

$$\begin{aligned} W &= \sum W_i \\ W_i &= Q_i/Q_{min} \end{aligned} \tag{3.20}$$

onde  $Q_{min}$  representa o quociente do mínimo caminho partindo da raiz até a espécie.

#### 3.7.1.4 Índices de Diversidade Filogenética Baseados em Caminho Mínimo

Os índices de diversidade filogenética baseados em caminhos mínimos possuem a propriedade de aferir o quão distante estão as espécies presentes na comunidade. Em outras palavras, quanto menor o caminho, menor é a quantidade de espécies e, conseqüentemente menor será a diversidade.

Uma medida quantitativa da diversidade filogenética ( $PD_{NODE}$ ) foi definida como o comprimento total mínimo de todos os ramos filogenéticos

necessários para medir um *táxon* em uma árvore filogenética. Valores de  $PD_{NODE}$  maiores podem ser esperados para corresponder a uma maior diversidade (FAITH, 1992). O índice  $PD_{NODE}$  é definido na Equação 3.21.

$$PD_{NODE} = \sum n_i \quad (3.21)$$

onde  $n_i$  representa o número de  $i$  nós no caminho mínimo de cada espécie presente na diversidade.

Diversidade filogenética incluindo ramos de base  $PD_{ROOT}$ , indicam o número de nós dentro do máximo caminho enraizado (RODRIGUES; GASTON, 2002). A Equação 3.22 define esse índice.

$$PD_{ROOT} = \sum n_{iROOT} \quad (3.22)$$

sendo  $n_{iROOT}$  o número de nós dentro do caminho. E por fim, tem-se a diversidade filogenética média ( $AvPD$ ), definida na Equação 3.23. Em que  $s$  indica o total de espécies.

$$AvPD = PD_{NODE}/s \quad (3.23)$$

### 3.8 Análise por Forma

A análise por forma é o estudo geométrico das formas que os objetos possuem. Especificamente um conjunto de medidas, que em sua essência caracterizam forma (circular, alongamento, por exemplo) e aparência dos objetos de uma imagem através do comportamento do contorno ou medidas extraídas através da área do objeto em estudo.

Partindo da premissa que os VOIs da TC tenham comportamento mais esférico e sabendo que regiões segmentadas a partir de técnicas de detecção perdem a maior parte das imperfeições do contorno, as medidas utilizadas neste trabalho irão aferir além das propriedades de forma do VOI, também as propriedades da forma do comportamento da textura da região intralesional.

### 3.8.1 Diagramas de Forma 2D

Diagramas de forma são representações de forma no plano euclidiano 2D ( $E^2$ ), usados para classificar conjuntos. Uma imagem pertencente a um conjunto é representada por um ponto dentro de um diagrama de forma cujas coordenadas são funções morfológicas, definidas a partir de inequações e funções geométricas (RIVOLLIER *et al.*, 2010b, 2010a).

Para representar uma forma, várias funções geométricas são determinadas, a fim de caracterizar os conjuntos. Elas estão relacionadas com as chamadas inequações geométricas que permitem definir funções morfológicas para criação dos diagramas da forma.

No processo de dedução dos diagramas de forma, seis funções geométricas foram usadas:  $A$ ,  $P$ ,  $r$ ,  $R$ ,  $w$ ,  $d$ , valores que denotam área, o perímetro, o raio inscrito e circunscrito, e o diâmetro mínimo e máximo de *Feret*, respectivamente. A Figura 3.7 apresenta um exemplo das referidas funções geométricas.

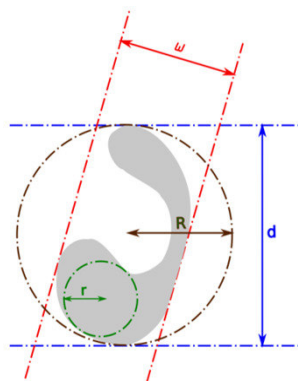


Figura 3.7: Funções geométricas: ( $r$ ) raio inscrito, ( $R$ ) raio circunscrito, mínimo diâmetro de *Feret* ( $w$ ) e máximo diâmetro de *Feret* ( $d$ ). Fonte (RIVOLLIER *et al.*, 2010b)

Para um conjunto em  $E^2$ , as relações entre essas funções geométricas são levadas pelas inequações geométricas listadas na segunda coluna da Tabela 3.2. Essas inequações estão ligadas às funções geométricas por pares e determinam os conjuntos que satisfazem as igualdades correspondentes.

As funções morfológicas são invariantes sob transformações semelhantes e são definidas como razões entre funções geométricas. Elas são normalizadas através de um valor constante (multiplicação escalar) que permite ter uma

proporção que varia no intervalo  $[0, 1]$ . Para cada função morfológica, o valor escalar depende diretamente da desigualdade associada (RIVOLLIER *et al.*, 2010a, 2010b). Essas funções morfológicas são referenciadas na terceira coluna da Tabela 3.2.

Tabela 3.2: Formas funcionais para os conjuntos em E2.

Funções Geométricas	Inequações Geométricas	Funções Morfológicas
$r, R$	$r \leq R$	$r/R$
$w, R$	$w \leq 2R$	$w/2R$
$A, R$	$A \leq \pi R^2$	$A/\pi R^2$
$d, R$	$d \leq 2R$	$d/2R$
$r, d$	$2r \leq d$	$2r/d$
$w, d$	$w \leq d$	$w/d$
$A, d$	$4A \leq \pi d^2$	$4A/\pi d^2$
$R, d$	$\sqrt{3}R \leq d$	$\sqrt{3}R/d$
$r, P$	$2\pi r \leq P$	$2\pi r/P$
$w, P$	$\pi w \leq P$	$\pi w/P$
$A, P$	$4\pi A \leq P^2$	$4\pi A/P^2$
$d, P$	$2d \leq P$	$2d/P$
$R, P$	$4R \leq P$	$4R/P$
$r, A$	$\pi r^2 \leq A$	$\pi r^2/A$
$r, w$	$2r \leq w$	$2r/w$

A partir dessas funções morfológicas descritas na Tabela 3.2, os 22 diagramas de forma podem ser definidos. Eles permitem representar a morfologia de qualquer conjunto no plano euclidiano 2D a partir de duas funções geométricas.

Para que qualquer conjunto seja representado, seis funções geométricas ( $\mathbf{A}$ ,  $\mathbf{P}$ ,  $\mathbf{r}$ ,  $\mathbf{R}$ ,  $\mathbf{w}$ ,  $\mathbf{d}$ ) e duas funções morfológicas ( $M1$  e  $M2$ ) são necessárias. Para qualquer diagrama de forma  $D$  representado no plano de domínio  $[0, 1]$ , em que as coordenadas do eixo são funções morfológicas  $M1$  e  $M2$ , qualquer conjunto 2D é mapeado em  $S$  por um ponto  $(x, y)$ . Matematicamente, um diagrama de forma de  $D$  é obtido a partir do seguinte mapeamento:

$$D : \begin{cases} K(\mathbb{E}^2) \rightarrow [0 - 1]^2 \\ S \mapsto x^2[0 - 1] \end{cases}$$

onde  $K(\mathbb{E}^2)$  denota os conjuntos do plano euclidiano 2D, usando funções morfológicas listadas na Tabela 3.2.



A ideia original exposta em Rivollier *et al.* (2010a, 2010b) descreve modelos matemáticos para criação de diagramas em planos euclidianos 2D, que são capazes de definir padrões de imagens que possuem semelhanças entre si. Foram realizadas adaptações para que seu uso se tornasse possível em objetos de três dimensões (3D). Tais adaptações referem-se às informações de área e perímetro em 2D que foram substituídas por equivalentes em 3D, que são: volume e área da superfície. Esses diagramas de forma não são criações diretas desta tese, pois os mesmos foram adaptados para uso nos métodos propostos.

### 3.8.1.1 Medidas de Proporção 3D

Para obter mais detalhes da forma do VOI, foram desenvolvidas medidas que visam aferir o quão esférico é o objeto analisado. Essas medidas combinam os parâmetros ( $\mathbf{A}$ ,  $\mathbf{P}$ ,  $\mathbf{r}$ ,  $\mathbf{R}$ ,  $w$ ,  $\mathbf{d}$ ) calculados para uso nos diagramas de forma (Seção 3.8.1).

**Convexidade** Essa medida tem como objetivo calcular o grau de esfericidade do objeto. Ela relaciona o maior diâmetro de Feret ( $d$ ) com o valor do diâmetro da menor esfera inscrita ( $r$ ). É definido pela Equação 3.24.

$$\text{convexidade} = \frac{d}{(r * r)} \quad (3.24)$$

**Proporção de Feret (PF)** Essa medida tem como objetivo calcular o grau de alongamento que o objeto possui. Assim, é calculada a razão entre o maior ( $d$ ) e o menor ( $w$ ) diâmetro de **Feret**, definido pela Equação 3.25. Com isso, é possível mensurar o quão alongado é o objeto.

$$PF = \frac{d}{w} \quad (3.25)$$

**Distância Média (DM)** Essa medida é calculada a partir do ponto central ( $x$ ,  $y$ ,  $z$ ) da menor esfera inscrita, a distância euclidiana para todos os pontos pertencentes ao objeto. Em seguida, é computada a média do somatório destes. Esta medida é definida pela Equação 3.26. Com isso, o quão mais próximo de 1 for o resultado final, mais esférico será o objeto.

$$MD = \frac{\sum_{i=1}^N \sqrt{(p - qi)^2}}{N} \quad (3.26)$$

onde  $p = (px, py, pz)$  representa o ponto central,  $qi = (qix, qiy, qiz)$  representa todos os demais pontos pertencentes ao objeto e  $N$  é o total de pontos que pertencem ao objeto.

**Regularidade esférica (RE)** Para o cálculo dessa medida, é realizada uma busca pelos raios das duas maiores esferas inscritas no objeto. Essa medida permite averiguar sobre a irregularidade esférica do objeto, ou seja, o quão mais próximo de 1 for o resultado da divisão, mais chance de ser irregular tem o objeto, este procedimento é definido pela Equação 3.27. O objetivo disso baseia-se no fato de que, quando calcula-se a primeira esfera inscrita dentro do objeto, acaba restando pouco espaço para calcular a segunda esfera, isto é, para objetos que possuem propriedades mais esféricas, o espaço restante após o cálculo da primeira esfera é menor, se comparado a objetos menos esféricos.

$$RE = \frac{r1}{r2} \quad (3.27)$$

onde  $r1$  e  $r2$  são, respectivamente, o maior e o menor raio das esferas inscritas ao objeto.

### 3.8.2 Abordagem Cilíndrica 3D

A análise da forma interna do VOI pode não ser suficiente para alcançar bons resultados. Encontrar uma maneira de obter informação com base na análise externa da forma do objeto é importante. Partindo desse princípio, acredita-se que a análise do VOI, por meio de sua projeção em um cilindro, é uma maneira viável de obter informações adicionais.

Optou-se por criar o cilindro apenas para delimitar uma região capaz de envolver todo o VOI, sem excesso ou perda de informações. Se outras formas geométricas fossem utilizadas para delimitar a região, tais como uma esfera, os resultados não seriam bons, pois, o raio formador da esfera poderia ser muito grande e iria causar a perda de informação nos VOIs que possuíssem grandes diferenças nos valores dos eixos  $X$ ,  $Y$  e  $Z$ . Esse problema não ocorre com um cilindro.

Antes de projetar o cilindro sobre o VOI, é necessário saber seu raio formador. Para tal, primeiramente é calculado o centro de massa da primeira

fatia do VOI. A partir desse, calculam-se quatro segmentos de reta, dois no *eixo-x* e dois no *eixo-y*. Em seguida, é escolhido o maior segmento. Esse procedimento é repetido para todas as fatias do VOI o maior de todos será usado como raio do cilindro. A Figura 3.8 mostra um exemplo do centro de massa, dos quatro segmentos e do maior segmento que será usado para formação do cilindro.

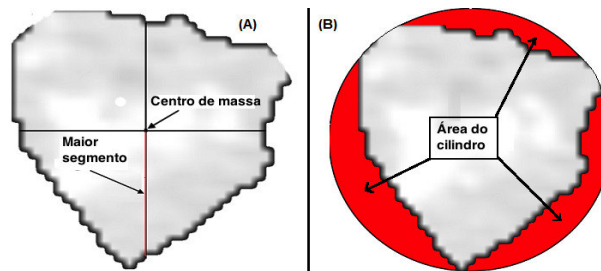


Figura 3.8: Exemplo de cálculo do raio para formar o cilindro. (A) exemplo de VOI com o centro de massa e segmentos de linha e (B) VOI com o cilindro criado.

Após o cálculo do raio, é criado o cilindro. A Figura 3.9 mostra o resultado do cilindro sobre um VOI. Na Figura 3.9A é mostrado um exemplo de VOI de nódulo composto por 3 fatias e na Figura 3.9B, um VOI de não-nódulo composto por 3 fatias. Em vermelho, tem-se a área do cilindro.

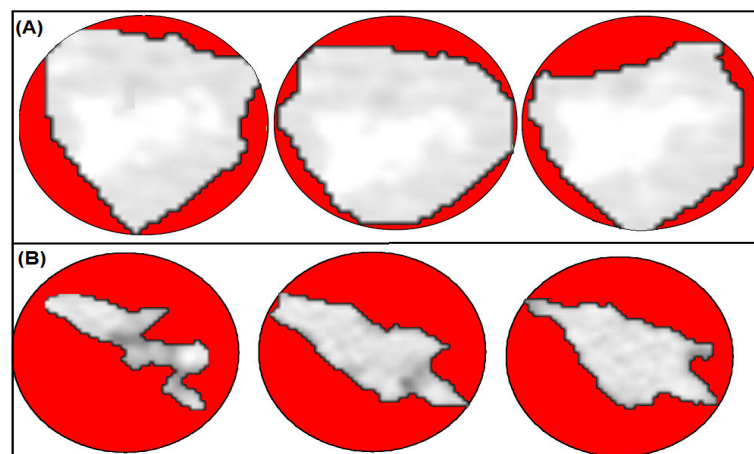


Figura 3.9: Exemplo de cilindros. A área do cilindro no VOI: (A) de nódulo e (B) de não-nódulo é mostrada em vermelho.

Depois de projetado o cilindro sobre o VOI, calculam-se duas medidas para obter informações de forma na região do cilindro (parte vermelha da Figura 3.9).

A primeira informação é calculada através da Equação 3.28. Ela permite verificar o valor médio do volume contido na região cilíndrica (vermelho) mostrada na Figura 3.9. O método de cálculo do volume que foi usado segue o mesmo apresentado nas obras de Gavrielides *et al.* (2009), Ko *et al.* (2003), Kuhnigk *et al.* (2006), Reeves *et al.* (2006), Udupa e Herman (1999). Essas obras estimam o volume usando apenas uma contagem de *voxel*. Através dessa informação, é possível analisar o tamanho dessa região, onde, quanto maior o valor, menor a probabilidade deste VOI possuir comportamento mais esférico.

$$meanvolume = \frac{nvoxel}{nnvoxel} \quad (3.28)$$

onde *nvoxel* representa o número de *voxels* do VOI e *nnvoxel* representa a quantidade de *voxels* que fazem parte da região cilíndrica. A Figura 3.10 mostra a região (azul) que será calculada em dois VOIs, sendo um de nódulo e outro de não-nódulo.

Além do cálculo do volume da região cilíndrica, é calculado também o raio da maior esfera inscrita contida na região cilíndrica para composição dos vetores de característica de cada VOI. Nesse caso, a esfera é o objeto mais adequado para este fim, pois permite a detecção de possíveis irregularidades no objeto. A Figura 3.10 mostra o resultado da esfera (vermelho) que foi encontrada em dois exemplos de VOI, um de nódulo (Figura 3.10A) e outro de não-nódulo (Figura 3.10B).

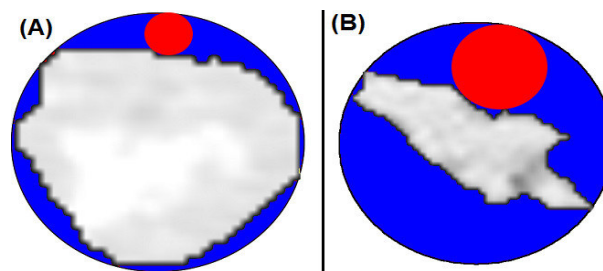


Figura 3.10: Exemplo do volume e da esfera calculados na região cilíndrica para: (A) VOI de nódulo e (B) VOI de não-nódulo.

Visando obter o maior número de informações que mensuram sobre o grau de esfericidade e alongamento do VOI, foram criados mais dois cilindros a partir do cilindro inicial, sendo estes reduzidos em 20% e 40% do valor original do raio.

Deste modo, tem-se um total de três cilindros com tamanhos diferentes para calcular o volume médio e o raio da maior esfera inscrita. A Figura 3.11

ilustra dois exemplos de VOI, um de nódulo e outro de não-nódulo, com seus respectivos cilindros, volume médio (em verde) e raios da maior esfera inscrita (em azul). Observa-se que o VOI de não-nódulo (Figura 3.11B) permite a extração do volume médio e do raio da maior esfera inscrita. Fato esse, não observado no VOI de nódulo (Figura 3.11A).

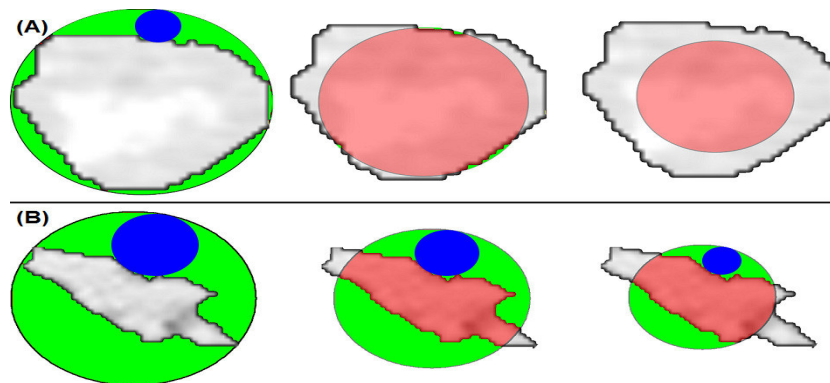


Figura 3.11: Exemplos dos três cilindros criados a partir de VOIs de (A) nódulo e (B) não-nódulo.

Vale ressaltar que os exemplos mostrados nesta seção são em 2D apenas para fins ilustrativos. Logo, toda a abordagem foi aplicada em objetos com três dimensões.

### 3.8.3 Funcionais de Minkowski

Os Funcionais de Minkowski (FMs) são um conjunto de ferramentas matemáticas extremamente robustas e eficientes, utilizadas para geometria integral (SANTALÓ, 2004), morfologia matemática (MECKE; STOYAN, 2002) e análise de imagens (DAVID *et al.*, 2011). Para a aplicação em imagens, seu uso está relacionado à análise quantitativa de padrões morfológicos (DAVID *et al.*, 2011; MECKE *et al.*, 1993).

Segundo o teorema de Hadwiger (HADWIGER, 1957), os FMs são aplicados a conjuntos que apresentam três propriedades principais: aditividade, invariância ao movimento e continuidade. Assim, em um espaço  $d$ -dimensional, as propriedades morfológicas globais de qualquer padrão podem ser completamente caracterizadas por  $d+1$  FMs. A descrição geométrica quantitativa tem como objetivo a redução da complexidade do objeto a um número limitado de medidas

relevantes (ROTH *et al.*, 2005). Essas devem satisfazer três requisitos, de maneira que os resultados obtidos para diferentes padrões e por diferentes observadores possam ser comparáveis. São eles:

**Aditividade:** O resultado obtido pela união de duas sub-regiões  $A$  e  $B$  deve ser o mesmo obtido pela soma dos resultados de duas sub-regiões individuais, menos a região de interseção entre as duas. Essa propriedade é importante porque o objeto não precisa ser analisado como um todo, mas por regiões. Essa propriedade é definida na Equação 3.29.

$$W(A \cup B) = W(A) + W(B) - W(A \cap B) \quad (3.29)$$

onde  $W(A)$  é uma determinada medida sobre o conjunto (objeto)  $A$ .

**Invariância ao movimento:** Os resultados obtidos devem ser constantes, mesmo que o objeto seja movido ou rotacionado, não dependendo assim da posição do observador.

**Continuidade:** Pequenas mudanças em um determinado objeto em análise podem gerar pequenas mudanças em suas medidas. No entanto, essa propriedade deve ser robusta para evitar que qualquer alteração, como ruídos, interfira nos resultados.

De acordo com o teorema de Hugo Hadwiger (HADWIGER, 1957), matematicamente, uma medida morfológica é definida como um funcional quando  $A$  e  $B$  são conjuntos (objetos) que possuem aditividade, continuidade e invariância ao movimento. Assim, uma grande quantidade de informações, obtidas por padrões da morfologia, pode ser comprimida em um número finito de parâmetros relevantes. Então, qualquer propriedade do tipo  $\varphi(A)$  (Equação 3.30) que depende da forma do objeto  $A$ , pode ser escrito como uma combinação linear de  $d+1$  FMs (ROTH *et al.*, 2005; MECKE; STOYAN, 2002).

$$\varphi(A) = \sum_{k=0}^d c_k W_k(A) \quad (3.30)$$

onde  $c_k$  são coeficientes reais que dependem da propriedade  $\varphi(A)$ , mas independem do objeto  $A$ .

**Propriedade 1:** A aditividade é uma boa propriedade matemática que permite com que os FMs possam ser calculados simplesmente baseados em uma

configuração de contagens.

Como visto em Michal *et al.* (2008), os FMs são definidos para um conjunto convexo e compacto  $A \subset R^3$  pelo teorema de *Steiner*. Assim, considerando que  $A \oplus B_r$  seja a dilatação do conjunto  $A$  em uma esfera fechada de raio  $r$  centralizada na origem, o volume  $V$  de  $A \oplus B_r$  pode ser escrito por uma função polinomial de  $r$ , como ilustrado na Equação 3.31:

$$V(A \oplus B_r) = \sum_{k=0}^3 W_k(A)r^3 \quad (3.31)$$

onde  $W_k$  é o  $k$ -ésimo Funcional de Minkowski. Por exemplo, considerando que  $C$  seja um cubo de lado  $a$ , então:

$$V(C \oplus B_r) = a^3 + 6a^2r + 3a\pi r^2 + \frac{4\pi}{3}r^3 \quad (3.32)$$

Com  $W_0(C) = a^3$ ,  $W_1(C) = 2a^2$ ,  $W_2(C) = a\pi$  e  $W_3(C) = 4\pi/3$ , é possível perceber a relação entre os FMs e os descritores, como volume ( $V$ ) (Equação 3.33), área da superfície ( $S$ ) (Equação 3.34), curvatura média ( $B$ ) (Equação 3.35) e número de Euler ( $X$ ) (Equação 3.36), respectivamente.

$$W_0(A) = V(A) \quad (3.33)$$

$$W_1(A) = \frac{1}{3}S(A) \quad (3.34)$$

$$W_2(A) = \frac{2}{3}\pi B(A) \quad (3.35)$$

$$W_3(A) = \frac{4}{3}\pi X(A) \quad (3.36)$$

Como há uma relação de proporção entre os FMs e os descritores de forma e conectividade ( $V, S, B$  e  $X$ ), é muito comum referenciá-los como Funcionais de Minkowski.

Em imagens 3D, são calculadas as medidas de Volume ( $V$ ), Área da Superfície ( $S$ ), Curvatura Média ( $B$ ) e Número de Euler 3D ( $X$ ). Esses parâmetros formam o chamado Funcionais de Minkowski (MECKE; STOYAN, 2002; DAVID *et al.*, 2011). O Volume está relacionado à massa que determinado objeto de

interesse possa apresentar. A Área da Superfície pode fornecer informações quanto à circularidade do objeto, isto é, valores baixos indicam uma maior proximidade com um objeto esférico. A Curvatura Média está relacionada à rugosidade da superfície, ou seja, valores altos indicam superfície mais rugosa e um valor menor pode indicar uma superfície mais suave. Por último, o Número de Euler refere-se à topologia da estrutura, isto é, relacionado à conectividade, indicando o número de estruturas conectadas menos o número de furos nessas estruturas.

Utilizando a **propriedade 1**, o cálculo dos FMs, em imagens 3D, é reduzido agora à basicamente uma contagem de *voxels*, isto é, calcular o número de cubos abertos  $n_3$ , faces abertas  $n_2$ , arestas abertas  $n_1$  e vértices  $n_0$ . De forma que na contagem, cada objeto (*voxel*) seja considerado apenas uma vez. Assim, tornando possível a adaptação para uso em nossos métodos.

Em estruturas discretas como imagens 3D, o cálculo dos FMs pode então ser representado na Equação 3.37.

$$\begin{aligned}
 V &= n_3, \\
 B &= \frac{3}{2}n_3 - n_2 + \frac{3}{2}n_1, \\
 S &= -6n_3 - 2n_2, \\
 X &= -n_3 + n_2 - n_1 + n_0
 \end{aligned}
 \tag{3.37}$$

### 3.8.4 Medidas de Distância

Na etapa de redução de falsos, positivos sabe-se que objetos mais alongados descaracterizam possíveis nódulos pulmonares. Desse modo, na etapa de diagnóstico, tem-se que o nódulo pulmonar, quando maligno, possui comportamento mais irregular e apresenta diversas formas, como por exemplo, um padrão espiculado ou com ramificações (Figura 3.12B e 3.12D), que podem expressar um crescimento desordenado das células cancerosas. Vale ressaltar que a especulação pode indicar também uma reação de fibrose causada pela presença do tumor ou pode ser uma área de atelectasia que indica uma vedação na entrada de ar, sendo estes também, motivos para delimitação de uma região interna ao nódulo. Já os nódulos benignos, o padrão de forma encontrado é o oposto (Figura 3.12A e 3.12C), neles se consegue obter, ao contrário do maligno, um padrão menos espiculado com menos ramificações (HANSELL *et al.*, 2008; GOULD *et al.*, 2001; FUJIMOTO; WISTUBA, 2014).



Essas irregularidades são de extrema importância para métodos que usam a forma do VOI para descrevê-lo. Entretanto, em alguns casos, esses padrões são perdidos ou reduzidos. Isto ocorre devido as próprias limitações operacionais dos especialistas, em outras palavras, ocorre uma variação na marcação intra-especialista e, por isso, a área demarcada pode conter excessos de tecido que não estão comprometidos pela lesão.

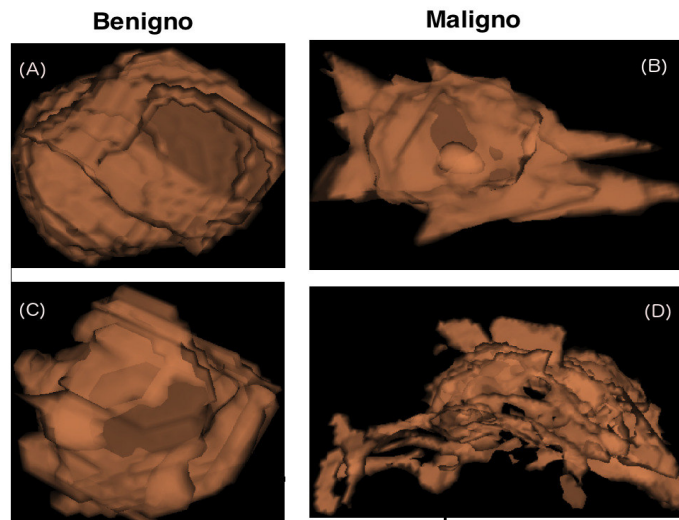


Figura 3.12: Exemplos de nódulos com padrões benigno e maligno.

As medidas de distância verificam o custo médio de distância necessário para que objetos alcancem, a partir de um determinado ponto de origem, outro ponto mais próximo.

A Equação 3.38 apresenta o custo médio de distância ( $Md1$ ) que o objeto interno consome para que, a partir de sua superfície, alcance a superfície de um objeto mais externo.

$$Md1 = \frac{\sum_{i=1}^n dist(A_i, B)}{n} \quad (3.38)$$

onde  $n$  representa o número de pontos da superfície do objeto interno,  $A$  representa o vetor de pontos do objeto interno e  $B$  representa o vetor de pontos do objeto externo, que para uso do método foram utilizados cilindros.

A segunda informação (Equação 3.39) verifica o custo médio de distância para se alcançar a superfície de um determinado objeto partindo do seu centro de massa ( $Md2$ ).

$$Md2 = \frac{\sum_{i=1}^n dist(P, A_i)}{n} \quad (3.39)$$

onde  $n$  representa o número de pontos da superfície do objeto,  $P$  representa o centro de massa e  $A$  representa o vetor de pontos da superfície do objeto.

Para ilustrar o comportamento das Equações 3.38 e 3.39, foi criada a Figura 3.13, que apresenta um exemplo em 2D. Na Figura 3.13A, tem-se em azul a representação do custo gasto por alguns pontos para se alcançar a borda do círculo e a Figura 3.13B ilustra o custo gasto para que se alcance a borda partindo do centro de massa ( $p$ ).

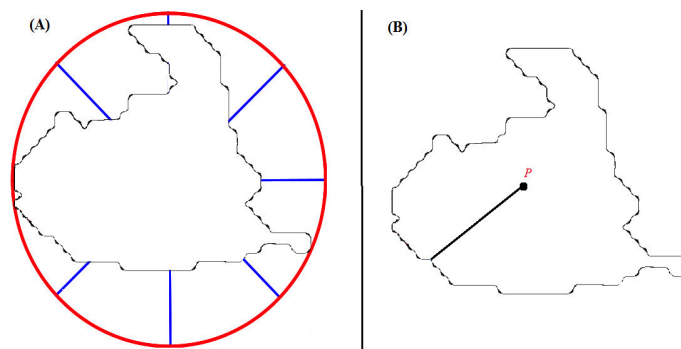


Figura 3.13: Exemplo em 2D dos descritores (A)  $Md1$  e (B)  $Md2$ .

### 3.8.5 Medidas Baseadas no Esqueleto

A definição de esqueleto diz que um ponto pertence ao esqueleto se ele for o centro de um círculo máximo, que toca a borda do objeto em pelo menos dois pontos distintos (GONZALEZ; WOODS, 2006). O algoritmo usado para extração do esqueleto consiste no método apresentado por Blum (1967).

Acredita-se que extrair medidas de um objeto partindo de seu eixo médio formador e que indique qualquer irregularidade que possua o objeto, irá fornecer informações essenciais. Para tal, propomos medidas que tomam como base informações extraídas a partir de seu esqueleto.

Do esqueleto, são extraídas as medidas descritas a seguir:

1. Volume do esqueleto.
2. Tamanho do eixo X.

3. Tamanho do eixo Y.
4. Tamanho do eixo Z.
5. Na Equação 3.40, tem-se a média da distância para a superfície ( $MT$ ).
6. A média dos diâmetros das esferas ( $MD$ ) é obtida através da Equação 3.41.

$$MT = \frac{\sum_{i=1}^n dist(A, B)}{n} \quad (3.40)$$

onde  $n$  representa o número de pontos do esqueleto,  $A$  representa o vetor de pontos do esqueleto e  $B$  representa o vetor de pontos do objeto que foi extraído o esqueleto.

Para cada ponto do esqueleto, são calculadas esferas, desde que estejam dentro do objeto formador do esqueleto. Na Figura 3.14, tem-se, apenas para fins ilustrativos, um exemplo usando um objeto 2D.

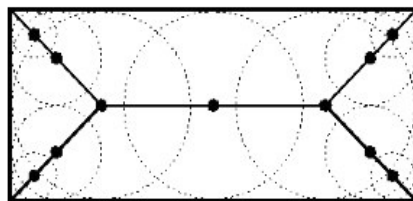


Figura 3.14: Exemplos de círculos que serão geradas para cada ponto do esqueleto. Fonte (SAMPAIO *et al.*, 2015).

$$MD = \frac{\sum_{i=1}^n diametro(A)}{n} \quad (3.41)$$

onde  $n$  representa o número de pontos do esqueleto e  $A$  representa o vetor de diâmetros calculados a partir de cada ponto do esqueleto.

### 3.8.6 Medidas Baseadas em Triangulação

Essas medidas têm o objetivo de averiguar as irregularidades do objeto em relação ao seu centro de massa e com base na sua superfície.

Quando os pontos da superfície mudam ao longo da forma, a área do triângulo formado por dois pontos da superfície e o centro de massa (Figura 3.15)

também muda. Isso forma uma função de área que pode ser explorada como representação da forma, onde  $S(n)$  representa a área entre os sucessivos pontos de fronteira  $P_i$ ,  $P_i + 3$  e  $G$  o centro de massa. Conforme a Equação 3.42, é feita a média de área do triângulo para cada objeto.

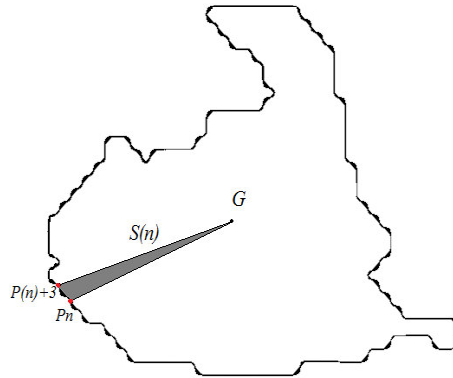


Figura 3.15: Área média do triângulo.

$$S(n) = \frac{\sum_{i=1}^n \text{area}(G, P_i, P_i + 3)}{n} \quad (3.42)$$

onde  $n$  representa o número de pontos da superfície do objeto.

A representação da área do triângulo (TAR) é calculada a partir da área dos triângulos formados pelos pontos na superfície da forma (ALAJLAN *et al.*, 2006, 2007). A curvatura do ponto do contorno  $(x_n, y_n)$  é medida usando a TAR, como se segue.

Para cada conjunto de cinco pontos consecutivos  $P_n - t_s$ ,  $x_n - t_s$ ,  $y_n - t_s$ ,  $P_n(x_n, y_n)$  e  $P_n + t_s$ ,  $x_n + t_s$ ,  $y_n + t_s$ , onde  $n \in [1, n]$  e  $t_s = 2$ . A área do triângulo formado por esses pontos é definida na Equação 3.43.

$$TAR(n, t_s) = \frac{1}{2} \begin{vmatrix} x_n - t_s & y_n - t_s & 1 \\ x_n & y_n & 1 \\ x_n + t_s & y_n + t_s & 1 \end{vmatrix} \quad (3.43)$$

O contorno é percorrido no sentido horário. Quando o valor da  $TAR$  (Equação 3.44) for positivo, estará indicando que os pontos são convexos; quando o valor da  $TAR$  for negativo, os pontos serão côncavos; e quando a  $TAR$  for igual a zero, indica uma reta. A Figura 3.16 demonstra esses três tipos de áreas triangulares.

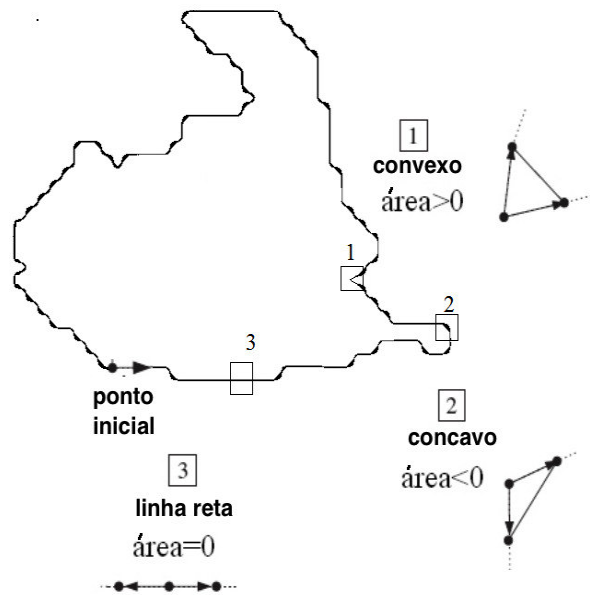


Figura 3.16: Exemplos de resultados para a TAR.

$$TAR_{mean} = \frac{\sum_{i=1}^n TAR_i}{n} \quad (3.44)$$

onde  $n$  representa o número de pontos da superfície do objeto.

### 3.8.7 Medidas Baseadas nos Diâmetros de Feret

Os diâmetros de *Feret* determinam a maior e a menor tangente de um objeto, ou seja, os tamanhos das tangentes em relação a cada eixo X, Y e Z.

As medidas baseadas nos diâmetros de *Feret* têm como objetivo calcular o grau de alongamento que um objeto possui. Assim, é calculada a razão entre o maior eixo Y ( $d$ ) e o menor eixo X ( $w$ ), definido pela Equação 3.48 e as proporções de cada eixo analisado do objeto interno (região intralesional) e objeto externo (VOI), conforme mostrado nas Equações 3.45, 3.46 e 3.47.

$$PX = \frac{x_{interno}}{x_{externo}} \quad (3.45)$$

$$PY = \frac{y_{interno}}{y_{externo}} \quad (3.46)$$

$$PZ = \frac{z^2}{(PX + PY)} \quad (3.47)$$

$$PF_m = \frac{PY}{PX} \quad (3.48)$$

Todas as técnicas propostas para caracterizar a forma do objeto dependem diretamente da técnica de segmentação/delimitação usada, uma vez que, esta influi diretamente na forma final do objeto. Para amenizar os possíveis erros que fatores como este podem provocar, nesta tese, é proposto uma delimitação interna ao VOI, de maneira automática, que visa isolar a região que teoricamente acredita-se ser a região lesionada.

### 3.9 Reconhecimento de Padrões

Na natureza, os padrões se manifestam de diversas maneiras, como sons, formas, imagens, cheiros e sabores.

Técnicas de Reconhecimento de Padrões são usadas para classificar ou descrever padrões ou objetos através de um conjunto de características extraídas. A todo instante, os seres humanos e também outros animais percebem e interagem com esses padrões com extrema naturalidade. Exemplos disso são as habilidades que o ser humano tem de diferenciar o som do motor de um automóvel do som de uma música ou ainda a habilidade que um animal selvagem tem de distinguir uma presa de um predador. A naturalidade inerente a essas habilidades faz com que o ser humano sequer imagine as complexidades cognitivas que estão por trás delas. Complexidades que se tornam evidentes quando se tenta reproduzir as tais habilidades artificialmente em um computador, o que há muito desafia a comunidade científica interessada no assunto (BISHOP, 2006).

O reconhecimento de padrão envolve duas tarefas: classificação, onde um conjunto de dados são divididos obedecendo alguma métrica formando classes ou grupos; e reconhecimento, em que um novo dado é reconhecido e atribuído em uma classe gerada. A classificação pode ser feita de duas formas: supervisionada e não supervisionada. No processo de classificação não supervisionada, tem-se um conjunto de dados que são subdivididos em grupos ou classes de acordo com algum critério de similaridade intra-classe. O processo de classificação supervisionada consiste em um processo prévio de treinamento de um classificador para o conhecimento dos padrões desejados.

Nesta tese, será usada para a etapa de classificação dos VOIs a Máquina de Vetor de Suporte.

### 3.9.1 Máquina de Vetor de Suporte

Máquina de Vetor de Suporte (MVS) é uma técnica de aprendizagem linear supervisionada, voltada para classificação e regressão de padrões. Uma característica especial dessa família de técnicas consiste no fato delas minimizarem o erro empírico de classificação, maximizando ao mesmo tempo a margem geométrica de erro. Com isso, essas técnicas são também conhecidas como classificadores de margem máxima (SCHWEIKERT *et al.*, 2009). O princípio básico é a construção de um hiperplano como superfície de decisão, cuja margem de separação entre as classes seja máxima (Figura 3.17).

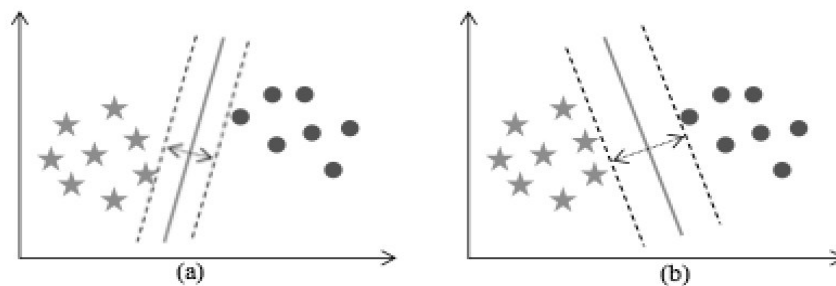


Figura 3.17: O conjunto de estrelas pertencentes a uma classe e o conjunto de círculos a outra classe. A margem de separação entre as classes é definida pelas retas tracejadas. Em (a) mostra um hiperplano para o conjunto de treinamento bidimensional com margem pequena e em (b) um hiperplano de margem máxima.

A técnica consiste no mapeamento de vetores de entrada em um espaço amostral com dimensão superior, por meio de construção do hiper-plano máximo de separação. Para isso, dois hiper-planos auxiliares são construídos de uma forma que o hiper-plano de separação resultante maximize a distância entre eles. É assumido que, quanto maior for a margem de distância entre os planos, menor é o erro de generalização do classificador (DUDA; HART, 1973).

#### 3.9.1.1 Funcionamento

Considerando  $X \in R_0 \subseteq R_n$  os vetores de entrada,  $y \in -1, +1$  os eixos e  $\varphi: R_0 \rightarrow F$  a função de mapeamento do espaço de entrada para o espaço de apresentação,

o algoritmo visa encontrar um hiper-plano  $(w, b)$  tal que o valor da Equação 3.49 seja máximo, sendo  $\gamma$  a margem, vetor  $w$  com a mesma dimensão de  $F$  e  $b$  um número real.

$$\gamma = \min_i y_i \{ \langle w, \phi(X_i) \rangle - b \} \quad (3.49)$$

A função de decisão correspondente é demonstrada na Equação 3.50.

$$f(X) = \text{sign}(\langle w, \phi(X) \rangle - b) \quad (3.50)$$

o mínimo da função  $f(X)$  ocorre quando a Equação 3.51.

$$w = \sum_i \alpha_i y_i \phi(X_i) \quad (3.51)$$

é satisfeita, sendo que são números reais positivos que maximizam a Equação 3.52:

$$\sum_i \alpha_i - \sum_{ij} \alpha_i \alpha_j y_i y_j \langle \phi(X_i), \phi(X_j) \rangle \quad (3.52)$$

e satisfazem a seguinte Equação 3.53:

$$\sum \alpha_i y_i = 0, \alpha_i > 0 \quad (3.53)$$

sendo que, dessa forma, é possível representar a função de decisão através da Equação 3.54.

$$f(X) = \text{sign} \left( \sum_i \alpha_i y_i \langle \phi(X_i), \phi(X) \rangle - b \right) \quad (3.54)$$

A partir da Equação 3.51, observa-se que  $\alpha_i$  associado ao ponto de treinamento  $X_i$ , expressa a importância daquele ponto no resultado final da função de decisão.

É importante observar que somente um subconjunto de pontos é associado a  $\alpha_i$  não-nulos. Tais pontos são denominados de vetores de suporte e são os pontos mais próximos ao hiper-plano de separação.

Dessa forma, a técnica possibilita a classificação de padrões de acesso, separando-os de acordo com os vetores de suporte do hiper-plano determinados. A técnica pode ser utilizada para predição de padrões numa sequência longa, conforme demonstrado em (SCHWEIKERT *et al.*, 2009).



### 3.9.2 Algoritmo evolutivo

Segundo Sampaio *et al.* (2015), os Algoritmos Evolutivos (AE) correspondem a uma classe de algoritmos utilizados para solucionar problemas de otimização através de modelos computacionais baseados na teoria da evolução das espécies, proposta por Charles Darwin, e nos princípios básicos da herança genética, descritos por Gregor Mendel (SCHUSTER, 2010).

Os Algoritmos Genéticos (AG) são um tipo clássico de AE amplamente utilizado. Porém, os AG sofrem do problema de convergência prematura. Para Sampaio *et al.* (2015), isto ocorre porque um bom indivíduo (ótimo local) quando encontrado, tende a trazer o processo de busca para o seu entorno, iniciando o processo de convergência para um ótimo local.

#### 3.9.2.1 Indivíduo $\alpha$

Na biologia, o indivíduo  $\alpha$  é o líder. Ele se destaca pela sua força, habilidade para caça, facilidade para tomar decisões, personalidade marcante e bravura. Na computação evolutiva, um grupo de indivíduos que se destaca é chamado de elite. A elite é constituída pelos melhores indivíduos de uma população numa determinada geração (SAMPAIO *et al.*, 2015).

Os indivíduos da elite não geram novos indivíduos, apenas fornecem material genético para a próxima geração. O elitismo garante que uma boa solução (cromossomo) não seja perdida na substituição da população na próxima geração (FILHO *et al.*, 2013).

#### 3.9.2.2 Metodologia de Evolução $\alpha$

A Metodologia de Evolução  $\alpha$  (ME $\alpha$ ) desenvolvida por Sampaio *et al.* (2015) baseia-se em alguns critérios biológicos importantes: presença do indivíduo  $\alpha$  (hierarquia social), pequenas populações e aumento da diversidade genética.

O trabalho desenvolvido por Sampaio *et al.* (2015) propõe uma ME $\alpha$  onde todas as operações genéticas ocorrem sobre o indivíduo  $\alpha$ . Essa metodologia trabalha com pequenas populações, de forma que qualquer ponto do espaço de busca tenha chance de ser percorrido, independente de qual geração um determinado cromossomo pertence (inicial, mediana ou final). Essa característica

lhe permite um aspecto dinâmico adaptativo às mudanças ambientais que possam ocorrer durante o processo de evolução.

A diferença da ME $\alpha$  desenvolvida por Sampaio *et al.* (2015) para as outras tradicionais é a etapa de criação da nova população. Primeiramente, deve-se definir o tamanho da população. Com o tamanho da população definida, deve-se criar aleatoriamente a população inicial e calcular o seu *fitness*. A evolução  $\alpha$  ocorre na criação da nova população (geração). Essa nova geração é formada pelo melhor indivíduo da geração anterior (indivíduo  $\alpha$ ) e de novos indivíduos criados de forma aleatória; e através de operações genéticas, entre essas, novos indivíduos e o indivíduo  $\alpha$ , exclusivamente. Repete-se o processo de criação dos novos indivíduos até que o tamanho mínimo da população seja superado. Calcula-se o *fitness* da nova população e verifica-se se o critério de parada foi alcançado. Em caso positivo, o indivíduo  $\alpha$  é a solução encontrada. Caso contrário, repete-se o processo a partir da criação da nova população. A Figura 3.18 ilustra o processo de escolha de evolução da ME $\alpha$ .

Em resumo, a ME $\alpha$  desenvolvida por Sampaio *et al.* (2015) segue as seguintes etapas:

1. Formar a base de análise (BA) a partir do conjunto de todos os vetores de características de cada VOI obtido na aquisição de imagem;
2. Cada elemento de A1 contém a posição de um vetor de características extraídas de um VOI da classe c1 na BA;
3. Cada elemento de A2 contém a posição de um vetor de características extraídas de um VOI da classe c2 na BA;
4. Os valores contidos na A1 e A2 são modificados pela mutação e pelos operadores genéticos de cruzamento, considerando que os valores não podem ser repetidos. Isso quer dizer que os elementos de A1 não podem estar presentes em A2 ou vice-versa. Os operadores genéticos de cruzamento e mutação não combinam dados de matrizes;
5. Os vetores de características selecionadas na A1 e A2 para cada indivíduo de uma geração são treinados pelo MVS (DUDA; HART, 1973). A MVS também é utilizada para testar os vetores de características da AB que não

foram selecionadas. Para medir o *fitness* de cada indivíduo, é utilizada a soma de especificidade, sensibilidade e precisão. Esse processo é repetido até que para 500 gerações consecutivas desse melhor indivíduo o *fitness* seja o mesmo;

6. Ao fim da evolução, todos os vetores de características, cujas posições estão contidas nas matrizes A1 e A2 do indivíduo com o melhor *fitness* na última geração, formam o melhor modelo de treinamento; e
7. O último passo é validar o modelo selecionado, através de uma classificação utilizando os nódulos restantes na base. E depois, é feito o cálculo da sensibilidade, especificidade e acurácia. Com isso, é possível medir a qualidade do modelo.

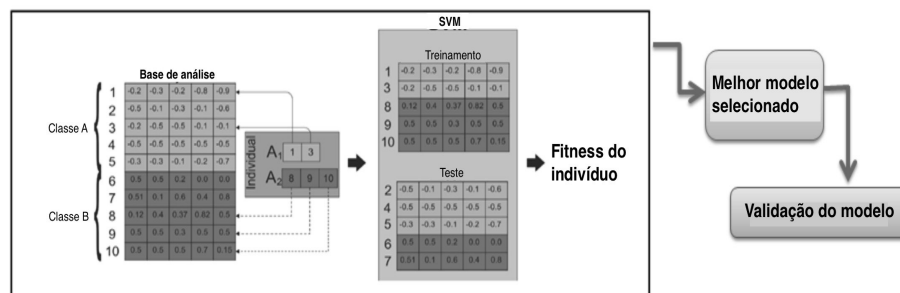


Figura 3.18: Ilustração da escolha do melhor modelo. Fonte (FILHO *et al.*, 2013).

Nesta tese, usou-se a  $ME\alpha$  para selecionar os melhores indivíduos para criação dos modelos de classificação usados pela MVS.

### 3.9.3 Análise Discriminante *Stepwise*

A Análise Discriminante é uma técnica comum no âmbito de aplicações envolvendo métodos de reconhecimento de padrões. Essa técnica utiliza informações das classes associadas a cada padrão para extrair linearmente os atributos mais discriminantes através da computação de uma combinação linear de  $m$  variáveis quantitativas que mais eficientemente separam grupos de amostras em um espaço  $m$ -dimensional, fazendo com que a razão da variância intra-classes e inter-classes seja maximizada (NASCIMENTO *et al.*, 2012).

O problema é então reduzido a achar um vetor que melhor represente esse conjunto de dados. A ideia básica que envolve a análise discriminante é determinar o quão as classes são distintas em relação à média de uma variável e, em seguida, usar essa mesma variável adequando-a a um grupo para a nova amostra.

Dois métodos computacionais podem ser utilizados para determinar uma função discriminante: o método simultâneo (direto) e o método *stepwise*.

A estimação simultânea envolve a computação da função discriminante e, como o próprio nome já ressalta, todas as variáveis independentes são consideradas juntas. Assim, a função discriminante é calculada baseada em todo conjunto de variáveis independentes, sem consideração do poder discriminatório de cada variável (HAIR *et al.*, 2006).

Na estimação *stepwise*, uma variável é selecionada conforme sua significância e, após cada etapa, as variáveis mais significativas são extraídas, compondo um conjunto de dados para investigação. O processo é iniciado escolhendo-se a melhor variável discriminatória. A variável inicial faz par então com uma das outras variáveis independentes, uma por vez, e a variável mais adequada, para melhorar o poder discriminatório da função em combinação com a primeira, é escolhida. As outras variáveis são escolhidas de maneira análoga. Após cada etapa de incorporação de uma variável, tem-se uma etapa em que variáveis previamente selecionadas podem ser descartadas. O procedimento é concluído quando nenhuma variável é incluída ou descartada (FILHO *et al.*, 2013).

Essa técnica será usada nesta tese após a extração das medidas de textura e forma para selecionar as melhores características.

### 3.9.4 Validação dos resultados

Após realizar as etapas de seleção dos indivíduos e a classificação, faz-se necessário validar os resultados obtidos. A validação do método é necessária para mostrar a sua eficácia na resolução do problema proposto.

Em problemas relacionados a classificação de dados, como neste caso, uma vez encontrados os candidatos a nódulos, pode-se classificá-los em nódulo e não-nódulo. Com isso, a validação se baseia em determinar o quão eficiente foi o teste de classificação. Nos problemas ligados à área de saúde, a estrutura básica dos testes de classificação é para determinar quão bem um teste discrimina

a presença ou ausência de uma doença (FILHO *et al.*, 2013).

Nesse tipo de problema, existe a presença de uma variável preditora (o resultado do teste) e uma variável resultante (a presença ou ausência da doença) a variável preditora do teste é geralmente dividida em duas (resultado negativo e positivo) e definindo os critérios de corte que confirmam ou não a presença da doença.

Quando se avalia um teste de classificação, quatro são as probabilidades possíveis de ocorrer:

1. O teste é positivo e a estrutura é realmente um nódulo - Verdadeiro Positivo (VP);
2. O teste é positivo, porém a estrutura não é um nódulo - Falso Positivo (FP);
3. O teste é negativo, quando na verdade a estrutura é um nódulo - Falso Negativo (FN); e
4. O teste é negativo e a estrutura também não é nódulo - Verdadeiro Negativo (VN).

Para avaliar o desempenho de modelos de classificação, geralmente utilizam-se algumas estatísticas descritivas, como Sensibilidade (Se), Especificidade (Es) e Acurácia (Ac). A sensibilidade (S) de um teste é definida pela proporção de nódulos com resultado do teste positivo, ou seja, é o número de nódulos detectados corretamente (VP) dividido pelo número total de nódulos existentes (dado por VP + FN). Tem a função de indicar a eficácia do método, Equação 3.55.

$$Se = \frac{VP}{VP + FN} \quad (3.55)$$

A especificidade (E) de um teste é a proporção de pessoas sem a doença que têm o teste negativo. Indica quão bom é o teste para identificar os indivíduos não doentes, ou seja, é o número de casos corretamente identificados como não-nódulos (VN) dividido pelo número total de casos que de fato não são nódulos (dado pela VN+ FP), Equação 3.56.

$$Es = \frac{VN}{VN + FP} \quad (3.56)$$

A acurácia de um teste refere-se à taxa de classificação correta. Ela é definida como a razão entre o número de casos corretamente classificados (dado pela  $VN + VP$ ) pelo número total de casos ( $VP + VN + FP + FN$ ), Equação 3.57.

$$Ac = \frac{VP + VN}{VP + VN + FN + FP} \quad (3.57)$$

Uma forma bastante aceita e utilizada de avaliar o desempenho quantitativo de uma determinada técnica ou modelo proposto pela comunidade científica são as chamadas curvas *Receiver Operating Characteristic* (ROC) (ERKEL; PATTYNAMA, 1998).

Essa forma de avaliação foi desenvolvida com bases estatísticas, em que a principal característica é a relação percentual de acertos. O ideal para um sistema é que a quantidade de acertos no diagnóstico, tanto em fatores VP quanto em VN, sejam máximas, conceitos assim definidos como sensibilidade e especificidade.

Do ponto de vista da análise das curvas ROC, isso significa que o ponto ótimo da curva é definido pelo extremo superior esquerdo que corresponde a máxima sensibilidade e mínimo número de FP, ou seja, máximo índice de VN. Consequentemente, quanto mais próximo de 1 (equivalente a 100%) a área sob a curva, melhor desempenho alcançado (ERKEL; PATTYNAMA, 1998). Na Figura 3.19, tem-se um exemplo da curva ROC.

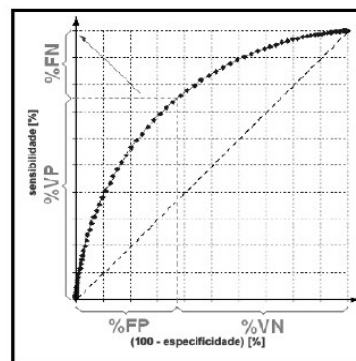


Figura 3.19: Curva ROC representando a relação entre a sensibilidade e a especificidade do classificador. Fonte (ERKEL; PATTYNAMA, 1998).

Como mostrado na Figura 3.19, cada ponto na curva ROC representa um par de valores (sensibilidade e especificidade) e a linha pontilhada diagonal um classificador que não consegue discriminar, devido ao percentual de VP ser igual

o percentual de FP.

Segundo Brown e Davis (2006), um importante índice para a análise da curva ROC é a Area Sob a Curva ROC<sup>6</sup>, sendo que um teste ideal é aquele cuja área sob a curva ROC é igual a 1 (discriminação ideal, no canto indicado pela seta azul). Porém, quando a curva ROC é a bissetriz, ou seja, área igual a 0,5 (sem discriminação, sob a linha diagonal tracejada), o teste não permite distinguir entre os grupos.

### 3.10 Considerações Finais

Este capítulo apresentou a fundamentação teórica usada no desenvolvimento desta tese. As teorias apresentadas são de extrema importância para a compreensão dos métodos propostos. Para tanto, foram apresentados conceitos sobre o câncer de pulmão, nódulo pulmonar solitário, tomografia computadorizada, arquivo e padrão DICOM, base LIDC-IDRI, sistemas auxiliados por computador para detecção e diagnóstico por imagem, processamento digital de imagem, algoritmo de Otsu, análise de textura, índices de diversidade filogenética, análise de forma, diagramas de forma, funcionais de Minkowski, medidas de proporção, abordagem cilíndrica, medidas baseadas: no esqueleto, em triangulação, nos diâmetros de Feret, reconhecimento de padrões, algoritmos genéticos e validação dos resultados.

Nos próximos capítulos serão abordados os materiais e os métodos aplicados na construção dos métodos para classificação, com o intuito de reduzir a quantidade de falsos positivos, isto é, classificar os VOIs em nódulo e não-nódulo e, por fim, os métodos de diagnóstico, que visam classificar os nódulos pulmonares em maligno ou benigno.

---

<sup>6</sup>do inglês *Area Under the ROC Curve*.

# Redução de Falsos Positivos

---

Neste capítulo, serão descritos os procedimentos adotados para o desenvolvimento dos métodos propostos para redução de falsos positivos. São apresentados dois métodos distintos, sendo um baseado na análise de textura e outro na análise de forma. Além desses, serão apresentados os resultados, as discussões e a comparação com outros trabalhos presentes na literatura.

## 4.1 Método

Em suma, a Figura 4.1 tem uma forma ampla das etapas seguidas pelos métodos propostos para redução de falsos positivos. Na primeira etapa, é feita a aquisição das imagens da base LIDC-IDRI (ARMATO *et al.*, 2011) e a composição das bases de imagens. Em seguida, é aplicada a extração e a seleção de características para os VOIs. Depois, realiza-se a seleção do melhor modelo e a classificação em nódulo e não-nódulo e, por fim, é feita a validação e análise dos resultados.

### 4.1.1 Aquisição de Imagens

Na base LIDC-IDRI, as imagens estão no formato DICOM e possuem até 16 bits por *voxel*. A base fornece um arquivo em formato XML (BRAY *et al.*, 1998) com a informação do contorno ao longo das fatias, além de algumas características, como esfericidade, textura, malignidade e etc. (a essas é indicado um valor entre 1 e 5) para aqueles nódulos pulmonares maiores que 3 mm. E apenas a informação sobre o centróide para aqueles inferiores a 3 mm. Um contorno é formado por



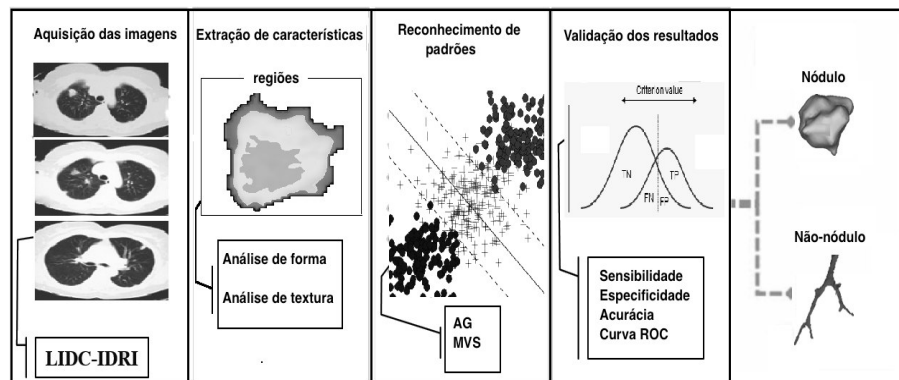


Figura 4.1: Esquema geral seguido pelos métodos propostos para redução de falsos positivos.

pares de pontos de coordenadas  $(x,y)$  que delimitam, em cada imagem do exame, a região onde o especialista identificou o nódulo.

O processo de anotação dos nódulos da base LIDC-IDRI foi feito por quatro especialistas 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.

A base LIDC-IDRI dispõe de 1.018 exames de TC, entretanto, alguns exames apresentaram dois fatores que tornaram inviável o uso no método, mais precisamente 185 exames. O primeiro fator está associado aos exames que não possuem nódulos maiores que 3 mm, pois, para esses nódulos, o arquivo com as marcações não dispõe dos contornos feitos pelos especialistas e nem com suas características, é encontrada apenas uma marcação com o centro de massa do nódulo. O segundo fator está relacionado a um erro de divergência de informações encontradas no arquivo de marcação dos exames, que não condizem com as encontradas no cabeçalho do DICOM daquele exame, tornando assim a marcação errônea e inválida. Assim, foram usados 833 exames para aplicação dos métodos.

A base LIDC-IDRI já é complexa, pois possui características, como aquisição de imagens com tomógrafos distintos e variações de contraste, que a torna de difícil análise. Além disso, vale ressaltar a grande quantidade de exames. Portanto, optou-se por selecionar um grande número de casos e dos mais variados

tipos (justapleural<sup>1</sup>, justavascular<sup>2</sup>, nódulos grandes, nódulos pequenos).

A amostra utilizada nesta tese tem dois tipos de VOIs, nódulos e não-nódulos. A seguir, disserta-se sobre a composição de cada grupo de VOI.

#### 4.1.1.1 VOIs de nódulo

Os nódulos foram automaticamente segmentados com base nas indicações previstas por especialistas do LIDC-IDRI. Foram utilizadas todas as regiões marcadas pelos especialistas como nódulos. Por exemplo, se a imagem possui quatro marcações, pode-se usá-las na formação da base como quatro nódulos distintos. A ideia é usar todas as recomendações do especialista. A Figura 4.2 mostra um exemplo de marcação por um especialista em uma imagem de TC.

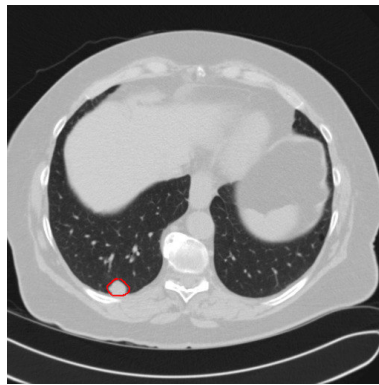


Figura 4.2: Exemplo de marcação em uma fatia TC.

#### 4.1.1.2 VOIs de não-nódulo

Uma vez que se tem a base de nódulos devidamente extraída e composta, é necessário adquirir os não-nódulos. Para isso, levou-se em consideração fatores que são importantes para compor o banco de imagens de não-nódulo. A seguir, enumeram-se os fatores:

1. O processo de extração manual é inviável porque se necessita de um grande número de não-nódulos para avaliar melhor os métodos propostos;

---

<sup>1</sup>Nódulos conectados a borda da pleura pulmonar.

<sup>2</sup>Nódulos conectados a vasos sanguíneos.

2. Uma forma automática de extração é necessária, sem qualquer indução ou intervenção humana/manual. Com isso, garante-se uma grande quantidade e com vários tipos de não-nódulos.

Com base nos itens acima, foi escolhida a metodologia automática de segmentação apresentada em Filho *et al.* (2013). Assim, usou-se os candidatos de não-nódulo que poderiam ser encontrados em um cenário real, já que um dos objetivos é integrar os métodos propostos a um sistema CAD.

A base de não-nódulos que será usada para teste dos métodos propostos foi baseada nas estruturas geradas na segmentação pelo trabalho de (FILHO *et al.*, 2013). Optou-se em usar essa estratégia porque: 1) não existe uma base de não-nódulos disponível; e 2) quer-se estruturas com características semelhantes ao nódulo. Esse trabalho é dividido em 3 fases: 1) extração do parênquima pulmonar e melhoramento das estruturas internas; 2) segmentação de estruturas semelhantes ao nódulo pulmonar; e 3) redução de falsos positivos. Na fase 2, foram segmentadas 18.292 estruturas, sendo nódulos e não-nódulos.

Com essas 18.292 estruturas, são selecionados os não-nódulos (17.742) para compor a base, excluindo os nódulos (550). Sabe-se o que são nódulos, baseado na marcação do especialista. Assim, se houver uma interseção de qualquer *voxel* entre alguma estrutura segmentada por Filho *et al.* (2013) e um nódulo marcado pelo especialista, tais estruturas são excluídos da base, pois é nódulo. Esse processo é repetido para todos os nódulos marcados pelo especialista e todas as 18.292 estruturas. Ao final do processo, tem-se apenas não-nódulos.

Em suma, foram extraídos 6.229 candidatos a nódulo a partir das marcações dos especialistas indicadas no XML; e 17.742 não-nódulos, baseado na metodologia de Filho *et al.* (2013).

Alguns exemplos de nódulo e não-nódulos são apresentados na Figura 4.3. A Figura 4.3A apresenta um caso de nódulo, a Figura 4.3B mostra um caso de não-nódulo e a Figura 4.3C, apesar de ter geometria semelhante a um nódulo, é apenas outra pequena estrutura. Casos como mostrados na Figura 4.3C são responsáveis por grande parte dos erros encontrados nas metodologias que utilizam apenas descritores de forma.

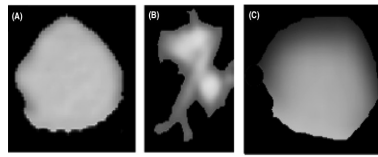


Figura 4.3: Três exemplos VOIs: (A) nódulo, (B) não-nódulo, (C) outra estrutura.

### 4.1.2 Regiões Internas

Análises baseadas na textura do nódulo pulmonar são amplamente exploradas, sejam elas na etapa de redução de falsos positivos (nódulo e não-nódulo) (FILHO *et al.*, 2013; MIYAKE *et al.*, 2009) ou para sugerir diagnóstico (maligno ou benigno) (NASCIMENTO *et al.*, 2012; OROZCO *et al.*, 2012).

Nesse contexto, enumeram-se alguns problemas que viabilizam e fundamentam a importância de uma análise por textura, que são: 1) em uma segmentação manual ou semi-automática, a delimitação do nódulo algumas vezes é muito maior que a área/volume do nódulo em si, ou seja, os especialistas, por algum motivo, delimitam uma região maior do que de fato seria necessária (FILHO *et al.*, 2013); 2) além disso, existem estudos que exploram a heterogeneidade intralesional do nódulo pulmonar (GERLINGER *et al.*, 2012), que podem fornecer informações histopatológicas do nódulo.

Baseado nas afirmações expostas no parágrafo anterior, é proposta uma nova segmentação interna ao VOI. Na Figura 4.4B, é mostrado um exemplo do excesso de tecido que pode ser delimitado no momento da segmentação manual ou semi-automática de um nódulo pulmonar (Figura 4.4A). E na Figura 4.4C pode ser observado uma região delimitada no interior da lesão indicando um padrão de distribuição de textura que pode ser explorado para ter as informações intralesional.

Para a tarefa de delimitação da região interna foi utilizado o algoritmo de Otsu (YANG *et al.*, 2012). A ideia básica do algoritmo de Otsu é aproximar o histograma de uma imagem por duas funções Gaussianas e escolher um limiar para minimizar a variância intra-classe, em outras palavras, o algoritmo Otsu encontra o limiar ótimo que separará as classes desejadas, isto é, como o objetivo é apenas separar a região acometida pelas células cancerosas de possíveis divergências intra-especialista, as quais possuem similaridades bem definidas, o método de Otsu

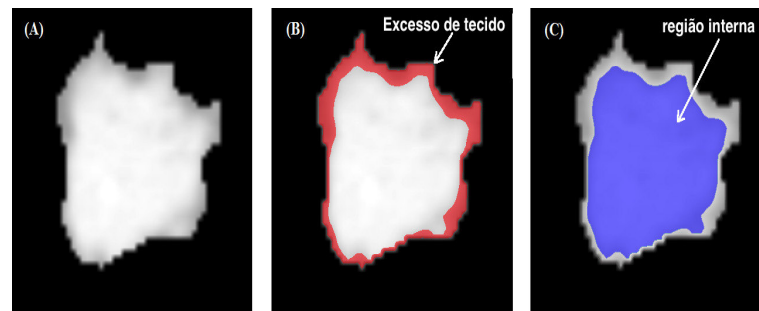


Figura 4.4: (A) imagem original, (B) imagem com o excesso de tecido, oriunda da segmentação e (C) imagem com a região interna à marcação original.

consegue separar bem as classes. A Figura 4.5 ilustra o comportamento dos níveis de cinza de um nódulo pulmonar, onde é possível observar a presença bem definida de duas classes, sendo elas, a que mais se assemelha ao tecido do parênquima pulmonar, na qual acredita-se ser o tecido intra-especialistas e os níveis de cinza, que de fato correspondem a lesão.

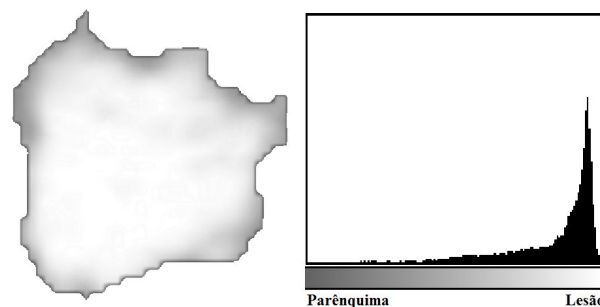


Figura 4.5: Exemplo de histograma que apresenta duas classes bem definidas.

Após a escolha do Otsu, é feita uma segmentação automática interna ao VOI, tornando possível uma nova perspectiva de análise das informações da textura intralesional, seja esta através da forma ou mesmo da textura. Para realizar tal procedimento, optou-se por gerar três regiões internas ao VOI com o algoritmo de Otsu (YANG *et al.*, 2012) (Seção 3.6.1). Foram usados dois limiares, sendo estes calculados automaticamente para cada VOI, dessa maneira, o VOI foi segmentado em três regiões. Na primeira região, objetivou-se delimitar a região mais periférica que contivesse possíveis excessos resultantes da segmentação, isto é, a análise dessa região tem como objetivo extrair informações dos tecidos periféricos que indicam padrões de comportamento de textura que ainda não foram ou estão

sendo acometidos pelas células cancerígenas.

Na segunda e na terceira região, tem-se as áreas que irão fornecer as informações intralesional do nódulo. O intuito em delimitar essas regiões está relacionado ao fato de tornar possível uma análise mais detalhada da região intralesional em si, de tal forma que os padrões de textura e forma em torno da região mais afetada mudem de acordo com o padrão de cada classe de lesão, isto é, maligno ou benigno. Na Figura 4.6 é mostrado um exemplo das três regiões que são formadas pelo algoritmo de Otsu.

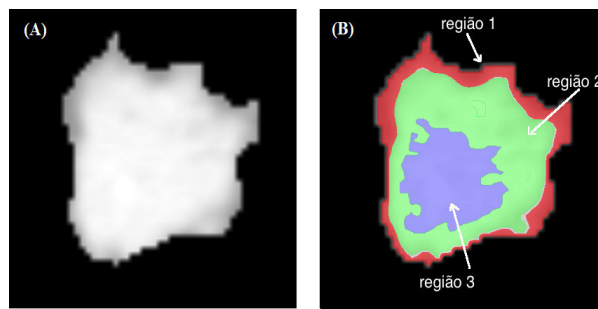


Figura 4.6: (A) imagem original, (B) 1 região periférica, 2 e 3 regiões intralesional.

Essas regiões serão usadas na etapa de extração de características que irão representar o VOI, para redução de falsos positivos nos métodos de forma e textura.

### 4.1.3 Extração de Características

Essa fase visa produzir medidas descritivas das imagens, as quais formarão os vetores de características que serão usados na etapa de classificação. Após a composição das bases de imagens e da estratégia de geração de regiões internas ao VOI, dar-se-á início a etapa de caracterização. O procedimento de extração de características adotado nesta tese será distribuído em duas vertentes, sendo uma baseada na textura e outra na forma.

#### 4.1.3.1 Extração de Características Baseada em Análise de Textura

A extração de informações da textura é de grande importância, pois, é através dessa que se torna capaz de mensurar propriedades muitas vezes imperceptíveis ao olho do especialista. Para descrição da textura, foram utilizados todos os índices

de diversidade filogenética apresentados na Seção 3.7.1. Esses índices são usados em conjunto com as árvores filogenéticas (Figura 4.8). Para montar e organizar a árvore, é necessário saber quais espécies estão presentes nos nódulos, sendo elas representadas pelas Unidades de Hounsfield (UH) presentes em cada nódulo pulmonar.

A seguir, descreve-se a fundamentação de como a árvore foi organizada e dos índices de diversidade usados.

**Árvore filogenética - dendrograma:** As árvores filogenéticas são usadas na Biologia para descrever as relações evolutivas entre as espécies. Nessas árvores, as folhas representam as espécies e os nós internos representam os antepassados comuns à espécie. Assim, é possível fazer uma conexão evolucionária entre as espécies estudadas. O dendrograma é uma representação gráfica que pode ser usada para descrever a relação filogenética entre espécies e seus antepassados (VANE-WRIGHT *et al.*, 1991).

Essas árvores permitem a extração de índices que medem a diversidade e o grau de parentesco entre as espécies (SILVA; BATALHA, 2006). A Figura 4.8 apresenta um exemplo de árvore filogenética, representado por um dendrograma. Nessa árvore, os nós folha são as espécies analisadas, os nós internos correspondem a algum ancestral comum e as arestas indicam a distância filogenética entre as espécies. Por meio de árvores filogenéticas, pode-se calcular índices de diversidade filogenética.

Para estabelecer a divisão das espécies no dendrograma, teve-se como objetivo inicial adotar uma estratégia baseada nas propriedades de textura presentes em cada nódulo, isto é, estabelecer uma forma de parentesco baseado na similaridade entre espécies. Para isso, foi adotada uma estratégia automática de segmentação baseada no algoritmo de Otsu (YANG *et al.*, 2012) (Seção 3.6.1). Optou-se por usar o algoritmo de Otsu porque, além de ser uma forma automática de segmentação não exige parâmetros iniciais.

A divisão realizada no VOI para montar o dendrograma deve assegurar um critério baseado na similaridade entre os *voxels*, isto é, no grau de parentesco entre as espécies. Como o objetivo do algoritmo de Otsu é encontrar um limiar ótimo que minimize a variância intra-classe através da aproximação do histograma

por duas funções Gaussianas, optou-se por usar o algoritmo de Otsu para realizar a divisão por similaridade entre as espécies. Vale ressaltar também que o custo computacional para realização da tarefa de divisão recursiva para montar o dendrograma é muito intenso e que, por ser um método de baixa complexidade, o algoritmo de Otsu respondeu bem a esse quesito.

Inicialmente, o algoritmo de Otsu segmenta o nódulo em duas ROIs com base em suas propriedades de textura, ou seja, a partir dos limiares gerados pelo algoritmo de Otsu tem-se duas ROIs, ROI esquerda, que contém todas as espécies (*voxels*) menores ou iguais ao limiar e ROI direita, que contém todas as espécies maiores que o limiar. A ROI esquerda é novamente segmentada pelo algoritmo de Otsu, gerando duas novas ROIs (direita e esquerda). Esse procedimento é repetido de forma recursiva para todas as ROIs que foram geradas, tanto esquerdas como direitas, até que o nível de espécies presentes em cada ROI seja igual a 1. A Figura 4.7 apresenta um exemplo dessa divisão.

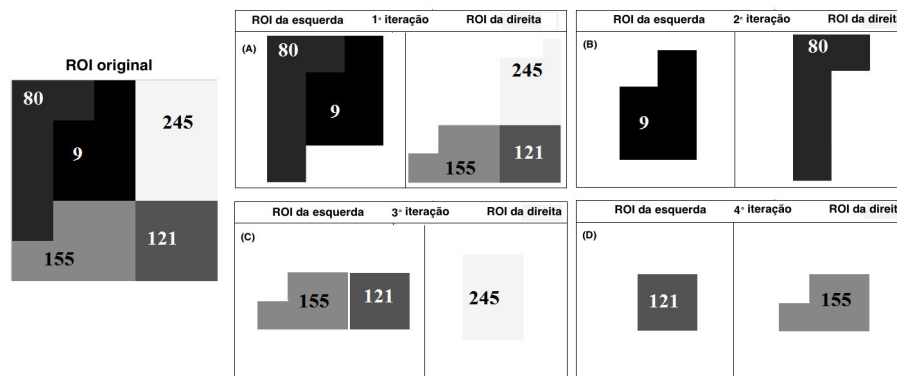


Figura 4.7: Quatro iterações realizadas pelo algoritmo de Otsu para separar a ROI original até restar apenas uma única espécie em cada folha. (A), tem-se a primeira iteração, na (B) a segunda iteração, na (C) e (D) a terceira e quarta iteração.

Com o procedimento de separação das espécies da ROI mostrado na Figura 4.7, tem-se o dendrograma (Figura 4.8) com suas espécies e suas características filogenéticas.

Com o dendrograma estabelecido, os índices de diversidade filogenética podem ser calculados para aferir as relações filogenéticas entre as espécies existentes na comunidade. Os índices serão extraídos tendo como base o VOI, tanto quanto as regiões que serão geradas a partir deste, seguindo o que foi descrito



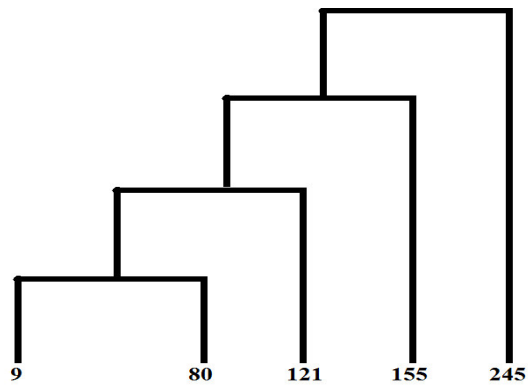


Figura 4.8: Dedrograma gerado a partir da ROI mostrada na Figura 4.7.

na Seção 4.1.2. Em resumo, na análise por textura, cada VOI será representado por 4 regiões, sendo elas: o VOI e as três regiões geradas a partir dele.

#### 4.1.3.2 Extração de Características Baseada em Análise de Forma

Assim como as características de textura, as de forma possuem um papel de extrema importância em métodos que possuem como objetivo a caracterização de VOIs (FILHO *et al.*, 2013; MIYAKE *et al.*, 2009).

Partindo do pressuposto que o VOI possui propriedades mais esféricas, quando se trata da redução dos falsos positivos, os descritores que serão utilizados para caracterizar os VOIs possuem na sua essência a capacidade de mensurar suas propriedades esféricas ou similares. Além disso, os descritores propostos serão capazes de aferir a forma da região intralesional, graças a estratégia de geração de regiões apresentada na Seção 4.1.2.

A partir das três regiões que serão geradas do VOI, apenas as duas regiões mais internas serão usadas pelos descritores de forma (Figura 4.6B, regiões 2 e 3). Essas regiões serão analisadas em conjunto, ou seja, serão consideradas como uma única região, pois se acredita que as regiões contêm a região intralesional, com isso, pode-se extrair medidas de forma mais precisas. Além disso, vale salientar que, usando também essa região, é possível minimizar os erros de classificação, uma vez que, além de analisar toda a forma do VOI, analisaria-se também a forma do núcleo intralesional. A região mais periférica (Figura 4.6B, região 1) não será usada, o que se dá pelo fato de que a mesma região contém os possíveis excessos de tecido demarcados pela marcação intra-especialista e não traria informações

de forma que fossem relevantes. Isso foi constatado através dos experimentos realizados.

#### 4.1.4 Reconhecimento de Padrões

Nesta etapa, verifica-se a capacidade de caracterização dos descritores propostos. Para tal, foi incluída, nos métodos propostos, uma etapa de reconhecimento de padrões, a qual é dividida em duas fases, uma de seleção do melhor modelo de treinamento e outra de classificação, usando o modelo selecionado.

##### 4.1.4.1 Seleção do Modelo

A criação de um modelo adequado para validação dos resultados é necessária. Para tal procedimento, foi aplicada, uma técnica evolutiva (Seção 3.9.2.2). Essa técnica fornece os melhores indivíduos da base de treinamento para criação desse modelo de classificação.

A  $ME\alpha$  evolui sempre baseado no indivíduo  $\alpha$ . Para uso nos métodos propostos, a evolução se deu da seguinte forma: foi estipulado primeiramente a quantidade de indivíduos que irá compor a população inicial que, de início, será o também indivíduo  $\alpha$ . Em seguida, é determinada a base que irá conter os indivíduos da população inicial e os demais que serão usados para realizar os cruzamentos. Esse cruzamento é a troca de material genético do indivíduo  $\alpha$  com população inicial, ou seja, essa troca é o processo em que o indivíduo  $\alpha$  fornece amostras que o compõe para a população inicial e vice-versa.

A evolução e o cruzamento realizado pela  $ME\alpha$  é baseada em uma classificação feita com o MVS retirando material genético do indivíduo  $\alpha$ , ou seja, quando se constitui a população inicial, é realizado um cruzamento com o indivíduo  $\alpha$  e, em seguida, é feita uma classificação onde essa população inicial servirá de modelo para classificação dos indivíduos restantes da amostra. No final da classificação, é calculado o *fitness* daquela população.

O valor do *fitness* é quem determina a qualidade daquela população para classificação dos demais indivíduos. O cálculo do *fitness* é feito baseado no resultado da classificação, isto é, a soma da sensibilidade, especificidade e acurácia, dessa forma, sempre que uma nova população apresenta um *fitness* melhor que o anterior, aquela população passará a compor o indivíduo  $\alpha$ . Assim, a  $ME\alpha$

evoluirá até que o critério de repetições seja atingido. Ao fim da evolução, têm-se os melhores indivíduos que serviram para criação do modelo de classificação.

#### 4.1.4.2 Seleção de características

A etapa de seleção de características é de grande importância, pois possui como objetivo selecionar, dentre todos os descritores, as características mais significativas da amostra. Foi utilizada a técnica Análise Discriminante *Stepwise* (Seção 3.9.3). Essa técnica faz a inclusão das variáveis independentes na função discriminante, uma por vez, com base em seu poder discriminatório. Dessa forma, espera-se que o conjunto de variáveis resultante contenha menos redundâncias.

A seleção das características é realizada antes da escolha do melhor modelo de treinamento. Desse modo, garante-se que todos os indivíduos presentes na base de imagens sejam levados em conta na etapa de seleção de características. Caso isso não seja feito dessa maneira, o resultado pode ser comprometido, pois, os VOIs que por ventura forem excluídos da base (na escolha do melhor modelo) podem ter impacto direto para um bom resultado na etapa de seleção das características.

#### 4.1.4.3 Classificação

Após o término da fase de extração de características e a seleção dos melhores indivíduos, os VOIs são classificados como nódulo e não-nódulo, no caso dos métodos de redução de falsos positivos e malignos ou benignos, caso o intuito do método seja o diagnóstico. Os vetores característicos são obtidos por meio das análises de forma e textura propostas. Esses valores são utilizados pelo classificador MVS com a *Radial Base Function* (RBF) (SCHÖLKOPF; SMOLA, 2002).

MVS é um poderoso algoritmo *state-of-the-art*, com fundamentos teóricos fortes baseados na teoria *Vapnik-Chervonenkis* (BLUMER *et al.*, 1989). MVS tem fortes propriedades de regularização. Regularização refere-se à generalização do modelo com novos dados. Essa característica foi a principal razão para a escolha desse classificador nesta tese. A precisão de um modelo MVS é altamente dependente da seleção dos parâmetros do *kernel*, tais como  $C$  e  $\lambda$ . Foi utilizado o software LIBSVM (CHANG; LIN, 2011) para estimar esses dois parâmetros.

### 4.1.5 Validação dos resultados

Para medir o desempenho dos métodos propostos, foram calculadas algumas estatísticas sobre os resultados dos testes.

Após a conclusão da fase de reconhecimento de padrões, é necessário validar e discutir os resultados. Esse método utiliza métricas comumente aplicadas em sistemas CAD/CADx para análise de sistemas baseados em processamento de imagem. As estatísticas calculadas foram: sensibilidade, especificidade e acurácia (Seção 3.9.4). Segundo Erkel e Pattynama (1998), outra maneira de medir o desempenho das técnicas de detecção e diagnóstico é a curva ROC, que indica a taxa de verdadeiros positivos (sensibilidade) como uma função da taxa de falsos positivos.

## 4.2 Resultados

Como mencionado anteriormente, cerca de 185 exames não foram utilizados nestes métodos devido a dois fatores que os tornaram inadequados. Um deles é o fato de que alguns desses exames não apresentam nódulos com diâmetro igual ou maior do que 3 mm e não incluem informações indicativas do grau de malignidade para o nódulo. O outro fator é a divergência de informações encontradas no arquivo de marcação do exame em relação a informação presente no cabeçalho DICOM do mesmo exame, tornando inviável o uso coerente das marcações (FILHO *et al.*, 2013). Dessa forma, foram utilizados 833 exames para treinamento e teste dos métodos propostos.

Tradicionalmente, as etapas de treinamento e teste dos classificadores é feita dividindo-se a base de teste e treino em percentuais que partem do pior cenário, isto é, de uma pequena proporção de treino e uma grande proporção de teste, até o melhor cenário, ou seja, grande parte para treino e uma pequena para testes. Alguns critérios devem ser destacados acerca dessa estratégia de testes. Geralmente, leva-se muito tempo na formação desse procedimento, assim, tem-se um grande espaço amostral para concepção dos modelos de treinamento, além disso, esse procedimento é repetido muitas vezes. Na maioria dos casos, os melhores resultados sempre estarão nas proporções que possuem um maior número de casos para treinamento. Além desses, tem-se ainda o desbalanceamento da

base, pois há uma grande disparidade de quantidade de dados entre as classes presentes no espaço amostral.

A  $ME\alpha$  foi empregado para gerar as bases de treinamento. A  $ME\alpha$  seleciona apenas os VOIs mais significativos de nódulo e não-nódulo para obter os melhores indivíduos e assim gerar o modelo de treinamento. É importante ressaltar a quantidade de *gerações* usadas até que a  $ME\alpha$  pare de evoluir, ou seja, caso o *fitness* se repita por 500 vezes (*gerações*) consecutivas, a  $ME\alpha$  chegará ao critério de parada. Além desses valores, os parâmetros utilizados pela MVS durante a seleção do melhor modelo foram os padrões, isto é,  $C = 1$  e  $\lambda = 0.5$ .

Com base nisso e nos resultados apresentados no trabalho de Filho *et al.* (2013), optou-se por empregar nos métodos propostos a  $ME\alpha$ . Com seu uso, consegue-se contornar os problemas expostos no parágrafo anterior, uma vez que, apenas os melhores indivíduos que representam as classes serão utilizados para criação do modelo de treinamento de forma balanceada. E por ser uma técnica de otimização, garante-se, através de seus critérios de evolução a participação de todos os indivíduos na busca pela solução ótima, até que o critério de parada seja alcançado.

A partir dos 833 exames e da metodologia Filho *et al.* (2013), foram extraídos 17.742 não-nódulos e da marcação do especialista foram extraídos 6.229 nódulos, totalizando 23.971 VOIs. Em seguida, foram selecionados de forma aleatória 80% (19.177) deles para formar a base de treinamento, garantindo a proporção para os nódulos e não-nódulos. Assim, os 20% (4.794) restantes foram usados para classificação, assim como para validar o modelo final selecionado pela  $ME\alpha$ .

Os 80% que constituem a base de treinamento foram submetidos a  $ME\alpha$  para seleção dos melhores indivíduos. A fim de selecionar apenas os indivíduos mais significativos e balancear a base de treinamento, deu-se como tamanho da população para a  $ME\alpha$  apenas 70% da base da classe com menor número (classe de nódulo). Dessa maneira, dos 80% originais da base de treinamento,  $ME\alpha$  irá selecionar apenas os 70% que melhor representem a classe nódulo e esta mesma quantidade seria usada para selecionar os da classe não-nódulo. O restante dos nódulos da base de treinamento foi utilizado pelo  $ME\alpha$  para validação do modelo. Esse procedimento foi aplicado a todos os experimentos realizados.

Os experimentos foram divididos em dois grupos. Primeiramente, são testados os descritores baseados em análise de textura e, em seguida, apresentam-se os resultados baseados em descritores de forma. E por último, serão apresentados os resultados para os experimentos utilizando os descritores de forma e textura em conjunto.

### 4.2.1 Resultados para os Descritores de Textura

Nesta seção, são apresentados os resultados para o classificador MVS usado nesta tese, com os descritores de textura baseado em índices de diversidade filogenética. Os experimentos foram realizados para cada grupo de índices, conforme descrito na Seção 3.7.1.

Na Tabela 4.1 e 4.2, são apresentados os resultados para redução de falsos positivos usando os índices de diversidade filogenética baseados na: riqueza de espécies (Seção 3.7.1.1); distância entre pares de espécies (Seção 3.7.1.2); topologia (Seção 3.7.1.3); e em caminho mínimo (Seção 3.7.1.4).

Na Tabela 4.1, são apresentados os resultados para os índices usando apenas como região de análise o VOI. E na Tabela 4.2, são apresentados os resultados para as análises usando 4 regiões, sendo o VOI e as 3 regiões, geradas como mostrado na Seção 4.1.2.

Primeiro, foram realizados testes usando como região de análise apenas os VOIs, ou seja, para o grupo de índices baseado na riqueza de espécies, apenas 2 atributos são extraídos para representar o VOI no vetor de características. Nos índices baseados na distância entre pares de espécies, são extraídas 5 características. Já nos índices baseados na topologia, são extraídos apenas 2 atributos. E nos índices baseados no caminho mínimo, 3 atributos compõem o vetor de atributos. E finalmente, faz-se um último teste usando os 12 índices em conjunto. A Tabela 4.1 apresenta os resultados para esses testes.

Os resultados da Tabela 4.1 apresentam resultados promissores. No melhor caso, tem-se quando combinados todos os índices para caracterização dos VOIs, alcançando uma acurácia de 97,24%, uma sensibilidade de 96,96, uma especificidade de 97,34% e uma área sob a curva ROC de 0,956. No pior caso, têm-se os índices baseados no caminho mínimo, com acurácia, sensibilidade e especificidade de 91,16%, 88,82% e 92,01%, respectivamente. E por fim, uma área

Tabela 4.1: Resultados para a redução de falsos positivos sem uso das regiões, usando análise de textura, baseado em índices de diversidade filogenética.

Índices baseados em:	Acurácia	Sensibilidade	Especificidade	ROC	Número de características
Riqueza de espécies	93,72	94,12	93,98	0,932	2
Distância entre pares de espécies	96,48	96,26	96,92	0,956	5
Topologia	95,88	93,54	95,88	0,926	2
Caminho mínimo	91,16	88,82	92,01	0,893	3
<b>Todos os índices</b>	<b>97,24</b>	<b>96,96</b>	<b>97,34</b>	<b>0,956</b>	<b>12</b>

sob a curva ROC de 0,893.

Seguindo com os experimentos, cada grupo de índice analisa 4 regiões para cada VOI. Dessa maneira, no grupo de índices baseado na riqueza de espécies, são extraídas 8 características por VOI (4 regiões x 2 índices). Nos índices baseados na distância entre pares, os vetores de características possuem 20 atributos (5 índices x 4 regiões). Já nos índices baseados na topologia, são extraídas 8 medidas (2 índices x 4 regiões). Em seguida, os índices baseados em caminho mínimo possuem 12 atributos (3 índices x 4 regiões) que caracterizam o VOI. E por fim, no último teste, tem-se a junção de todos os índices, assim, cada VOI foi descrito no vetor por 48 características. Na Tabela 4.2, têm-se os resultados.

Tabela 4.2: Resultados para a redução de falsos positivos com as regiões, usando análise de textura, baseado em índices de diversidade filogenética.

Índices baseados em:	Acurácia	Sensibilidade	Especificidade	ROC	Número de características
Riqueza de espécies	98,38	98,32	98,39	0,976	8
Distância entre pares de espécies	99,57	99,24	99,69	0,981	20
Topologia	99,28	99,40	99,23	0,979	8
Caminho mínimo	98,97	96,12	99,66	0,946	12
<b>Todos os índices</b>	<b>99,57</b>	<b>99,45</b>	<b>99,61</b>	<b>0,992</b>	<b>48</b>

Como visto na Tabela 4.2, o uso das regiões proporciona uma melhora nos resultados, com isso, foi possível obter resultados mais promissores. No melhor caso, a junção de todos os índices proporciona resultados mais favoráveis, pois, foi possível alcançando uma acurácia de 99,57%, uma sensibilidade de 99,45%, uma especificidade de 99,61% e uma área sob a curva ROC de 0.986. No pior caso, destacam-se novamente os índices baseados no caminho mínimo, com acurácia,

sensibilidade e especificidade de 98,97%, 96,12% e 99,66%, respectivamente. E por fim, uma área sob a curva ROC de 0,946.

Em suma, acredita-se que os índices baseados no caminho mínimo não demonstram um bom resultado porque a informação extraída relacionada ao grau de parentesco (diversidade) entre as espécies não consegue ser bem evidenciada, uma vez que esse tem seu comportamento central baseado na raiz da árvore filogenética.

Como pode ser observado, à medida que a quantidade de características aumentam os valores também melhoram, acredita-se que esse fato ocorre porque os índices se complementam, isto é, um índice pode destacar uma propriedade que outro não consegue.

### 4.2.2 Resultados para os Descritores de Forma

Nesta seção, serão apresentados os resultados para a redução de falsos positivos usando os descritores de forma apresentados na Seção 3.8. Os resultados seguirão a mesma lógica mostrada na Seção 4.2.1, isto é, foram realizados testes com cada grupo de descritores usando para análise apenas o VOI e, em seguida, os resultados usando como região além do VOI, a região mais interna obtida pela segmentação apresentada na Seção 4.1.2, como explicado na Seção 4.1.3.2.

Os resultados apresentados na Tabela 4.3 não são muito promissores se comparados aos da Tabela 4.2, isto se dá pelo fato da complexidade que é uma análise de forma, pois, observou-se que as formas obtidas nos VOIs, tanto de nódulo quanto de não-nódulo, eram por vezes semelhantes, tornando assim a análise por forma propícia a erros. No melhor caso, tem-se quando são usados todos os descritores para caracterização do VOI, perfazendo uma taxa de acerto promissora, se comparado aos demais resultados da tabela. Assim, foi possível atingir uma acurácia de 94,52%, sensibilidade de 92,30%, especificidade de 96,13% e uma curva ROC de 0,918. No pior, caso pode-se destacar os descritores baseados nos diagramas de forma.

Já na Tabela 4.4, os resultados apresentam-se mais promissores, uma vez que, combinando a região intralesional e o VOI, a capacidade de caracterização dos descritores aumenta. Dessa maneira, o melhor cenário está quando combinados todos os descritores, alcançando uma média de acerto geral de 98,15%,



Tabela 4.3: Resultados dos descritores de forma na redução de falsos positivos analisando apenas o VOI.

Descritor	Acurácia	Sensibilidade	Especificidade	ROC	Número de características
Diagramas de forma	84,91	75,40	88,44	0,810	26
Abordagem cilíndrica	85,26	76,54	89,18	0,814	6
Funcionais de Minkowski	83,87	79,74	85,50	0,804	4
Medidas de distância	87,23	86,45	87,34	0,833	2
Medidas do esqueleto	87,01	79,76	90,45	0,795	6
Medidas de triangulação	83,94	79,81	85,13	0,819	2
Medidas baseadas nos diâmetros de Feret	86,74	87,45	86,14	0,846	4
<b>Todos os descritores</b>	<b>94,52</b>	<b>92,30</b>	<b>96,13</b>	<b>0,918</b>	<b>50</b>

Tabela 4.4: Resultados dos descritores de forma para redução de falsos positivos analisando o VOI e a região intralesional.

Descritor	Acurácia	Sensibilidade	Especificidade	ROC	Número de características
Diagramas de forma	94,57	91,54	95,58	0,908	52
Abordagem cilíndrica	94,57	90,78	95,92	0,904	12
Funcionais de Minkowski	94,70	90,75	96,09	0,902	8
Medidas de distância	94,80	93,03	95,43	0,902	4
Medidas do esqueleto	90,06	88,28	90,80	0,882	12
Medidas de triangulação	92,77	89,75	95,10	0,898	4
Medidas baseadas nos diâmetros de Feret	94,72	91,36	95,91	0,910	8
<b>Todos os descritores</b>	<b>98,15</b>	<b>96,75</b>	<b>98,87</b>	<b>0,967</b>	<b>100</b>

sensibilidade e especificidade de 96,75% e 98,87%, respectivamente, com uma área sob a curva ROC de 0,967. No pior caso, pode-se destacar os descritores baseados no esqueleto, perfazendo uma acurácia de 90,06% e uma curva ROC de 0,882. Isso se deu através do processo de formação do esqueleto de alguns VOIs de não-nódulos possuírem forma bem próxima a de VOIs de nódulos, com isso, acometendo nos erros.

### 4.2.3 Resultados com todos os descritores e seleção de características

Como os descritores de forma e textura na teoria se complementam, são realizados dois últimos experimentos. No primeiro experimento, todas as características extraídas, tanto de forma quanto de textura, são usadas para descrever o VOI. Os resultados foram bem promissores, alcançando uma acurácia de 99,41%, uma sensibilidade de 99,59%, uma especificidade de 99,55% e curva ROC de 0,993. Se comparado com o melhor resultado apresentado na Tabela 4.2, o resultado ficou bem próximo, apresentando uma leve melhora na sensibilidade, mas diminuindo um pouco a especificidade.

Quando se analisa a quantidade de características, tem-se um número bem expressivo, isto é, um total de 148 medidas (48 de textura e 100 de forma). Baseado nisso, fez-se um último experimento, no qual foi aplicada a Análise Discriminante *Stepwise* (Seção 3.9.3) para selecionar as características mais significativas.

Após a seleção, resultaram apenas 18 características, sendo elas: 2 dos índices baseados na riqueza de espécies; 4 dos índices baseada na distância entre pares de espécies; 3 dos índices baseados na topologia; 3 das medidas de distância; 2 da abordagem cilíndrica; 2 dos diagramas de forma; 2 dos diâmetros de Feret. Vale ressaltar que as características foram selecionadas com todos os indivíduos da base, isto é, antes de selecionar o melhor modelo.

Os resultados não apresentam melhorias, entretanto, mostram-se promissores tanto quanto o do melhor resultado apresentado na Tabela 4.2. Em sumo, tem-se uma sensibilidade de 99,12%, uma especificidade de 99,05%, uma acurácia de 99,09% e uma curva ROC de 0,990. Vale ressaltar a quantidade de características usadas, pois, conseguiram-se resultados bem próximos ao melhor

caso, mas utilizando um número de medidas bem menor.

Finalmente, na Figura 4.9, são apresentadas as curvas ROC para os melhores resultados da análise de: textura, forma, todas as medidas em conjunto e após a seleção de características, para métodos de redução de falsos positivos.

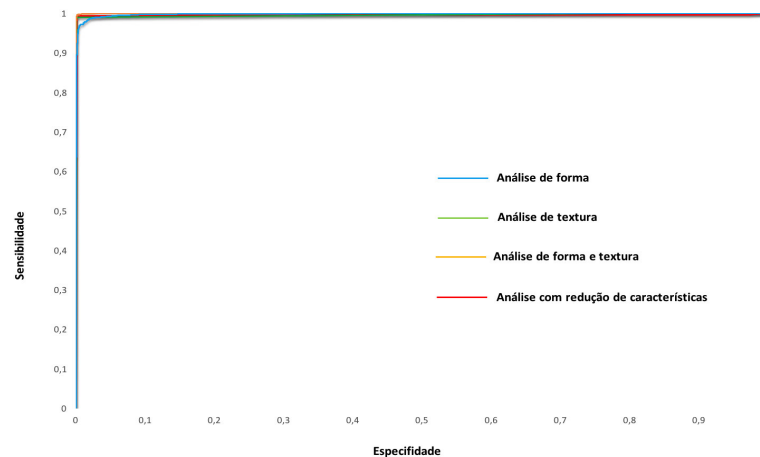


Figura 4.9: Resultados para as curvas ROC dos métodos de redução de falsos positivos.

### 4.3 Discussão

Partindo dos resultados apresentados nas Tabelas 4.1 e 4.2, os métodos propostos para a redução de falsos positivos, baseada em análise de textura, demonstram-se bastante promissoras. As análises de textura através dos índices de diversidade filogenética combinados com árvores filogenéticas conseguiram resultados muito expressivos em todas as análises com os grupos de índices.

Dentre todos os índices, o grupo de índices baseado na distância entre pares de espécies apresentou os melhores resultados. Acredita-se que isso ocorra devido a avaliação da diversidade presente na comunidade, analisando precisamente a relação existente entre as espécies, além disso, esse grupo de índices possui um maior número de descritores, o que também influenciou nos resultados, pois mais detalhes puderam ser avaliados.

Acreditando no potencial dos descritores, é necessário descrever sobre o experimento realizado com todos os índices em conjunto. A premissa básica para realização desses índices se dá pelo fato que um índice possa identificar alguma

propriedade da textura do VOI que outro não consiga. Partindo dessa premissa, observa-se que em todos os casos dos resultados apresentados nas Tabelas 4.1 e 4.2 a combinação dos índices mostra-se mais eficiente, isto é, houve uma melhora importante nos resultados.

Além dos motivos expostos no parágrafo anterior, a proposta de criação das regiões internas ao VOI proporcionou uma melhora nos resultados. Acredita-se que essa melhora se deu pelo fato da análise da textura ter sido feita em partes específicas do VOI, dessa maneira, conseguindo destacar propriedades de textura não apenas da região intralesional, mas de suas áreas periféricas. Em outras palavras, o uso das regiões tornou possível a análise isolada das regiões partindo do núcleo da lesão, até as regiões mais periféricas, que teoricamente possuem propriedades de textura distintas aos do núcleo pelo fato de ainda estarem sendo acometidas pelas células cancerosas.

Ainda analisando os resultados para a redução de falsos positivos, têm-se as Tabelas 4.3 e 4.4 que demonstram os resultados para os experimentos usando apenas os descritores de forma propostos. A análise utilizando apenas descritores de forma não é uma tarefa simples, pois a geometria do VOI algumas vezes pode acometer erros de classificação. É comum encontrar na literatura sistemas CAD que utilizam técnicas de segmentação que geram regiões suspeitas por meio do grau de semelhança entre seus níveis de cinza, isto é, regiões com textura muito semelhante. Dessa forma, faz-se necessário ter uma alternativa de análise. Com base nisso, foram desenvolvidos novos descritores e adaptados outros, para caracterização dos VOI com intuito de reduzir quantidade de falsos positivos.

Apesar dos diagramas de forma possuírem uma grande quantidade de características, seus resultados não são tão eficientes quanto os demais descritores, isto mostra que uma grande quantidade de descritores não implica em bons resultados. Acredita-se que esses resultados relacionam-se com a pouca capacidade de detalhes extraídos a partir dessas técnicas, uma vez que a anatomia do VOI não favoreceu o bom desempenho para esse conjunto de descritores.

Os resultados demonstram-se promissores, pois alcançaram valores expressivos para a dedução de falsos positivos. Para os resultados da Tabela 4.3 a eficiência dos descritores utilizados em conjunto demonstra a efetividade dos métodos propostos. Na Tabela 4.4, é perceptível a melhoria dos resultados quando

combinados os descritores com a delimitação da região intralesional do VOI. Esses resultados asseguram que o padrão de forma obtido a partir da região interna ao VOI é de extrema importância na caracterização do mesmo. Isso destaca mais uma vez a importância de analisar o interior do VOI.

É importante mencionar que nossa abordagem baseada em cilindro pode falhar em casos em que o objeto analisado seja um vaso sanguíneo, pois, caso este esteja orientado na direção do eixo Z, para nosso método, esse vaso será um cilindro perfeito. Entretanto, esse erro é contornado mediante a análise baseada nos descritores de textura.

Na literatura, é comum encontrar trabalhos que utilizem descritores de forma e textura para realizar a tarefa de caracterização dos seus indivíduos. Os nossos resultados quando combinados todos os descritores, foram bem satisfatórios, pois alcançaram valores acima dos 99% de acerto. Isso mostra a eficiência dos descritores propostos em diversos tipos de combinação de análises.

Cabe ressaltar também a eficiência dos descritores após a seleção dos melhores atributos, pois foi possível alcançar valores expressivos acima dos 99% em todas as métricas de avaliação com um conjunto de apenas 18 atributos.

Em resumo, ambos os métodos utilizadas para a redução de falsos positivos em VOIs de TC demonstraram ser eficientes na tarefa de classificação em nódulo e não-nódulo. Os resultados obtidos são muito expressivos, uma vez que alcançaram, no melhor caso, taxas de acerto entre 98% e 99%, para forma e textura, respectivamente. E acima dos 99% para todos os descritores em conjunto. Cabe destacar ainda que os métodos baseados na textura combinado com a delimitação das regiões internas ao VOI apresentaram resultados pouco acima dos resultados com os descritores de forma.

Além da base LIDC-IDRI ser uma base complexa e de grande diversidade de exames, ela também conta apenas com exames convencionais de TC, isto é, não usam elementos químicos para elevação do contraste das estruturas para facilitar na redução de falsos positivos, isso é importante para mostrar a eficiência dos métodos propostos, pois alcançam bons resultados sem a exposição a uma maior radiação que possam por em risco a saúde do paciente e do especialista.

Os métodos propostos demonstram ser muito promissoras se comparados aos demais encontrados na literatura, uma vez que os resultados obtidos foram

bem acima do padrão *gold*<sup>3</sup> (estimado em média a 85%) (BRAZ JUNIOR *et al.*, 2013), superando os 99% de acerto total. Além disso, esses resultados asseguram a importância que tem a análise da região intralesional, uma vez que, em todos os casos, proporcionaram-se bons resultados, conseguindo melhorar os resultados de forma considerável.

E por fim, os bons resultados asseguram a efetividade da ME $\alpha$  em selecionar os melhores indivíduos para composição do melhor modelo.

## 4.4 Comparação com outros Trabalhos

Nesta seção, serão comparados os métodos propostos com outros trabalhos da mesma natureza. A comparação com outros trabalhos na área é uma tarefa difícil, porque nenhum dos trabalhos relatados nesta tese forneceu os exames utilizados. A única informação disponibilizada é o banco de dados utilizado. Assim, não foi possível realizar uma avaliação rigorosa de nossos métodos em relação a outras obras.

O objetivo das Tabelas 4.5 e 4.6 é fornecer uma visão geral (base de exames usado, métricas de avaliação de desempenho) dos resultados encontrados nos trabalhos relacionados com as nossas propostas. Assim, pretende-se mostrar que os métodos propostos são promissoras, uma vez que, em comparação com outras obras, alcançamos resultados acima de 98% para redução de falsos positivos.

A comparação foi realizada de duas maneiras. Primeiramente, na Tabela 4.5, comparam-se os métodos propostos com trabalhos que utilizaram apenas medidas de textura para composição de seus vetores de características. E na Tabela 4.6, é apresentada a comparação dos nossos métodos com outros trabalhos que utilizaram medidas de forma e textura para caracterizarem seus nódulos.

É importante mencionar que as comparações realizadas na Tabela 4.5 são apenas de caráter expositivo/comparativo, uma vez que, na maioria dos casos as bases usadas divergem, os métodos usados pelos autores não estão acessíveis ou

---

<sup>3</sup>Valor de referência utilizado como parâmetro para estabelecer uma meta de acerto, baseada nas experiências dos especialistas.

os parâmetros utilizados não são fornecidos. Tudo isso impacta diretamente para uma comparação fidedigna.

Tabela 4.5: Comparação com trabalhos que usam apenas descritores de forma para redução de falsos positivos. Legenda: acurácia (Ac), sensibilidade (Se), especificidade (Es) e área sob a curva ROC.

Trabalho	Base	Amostra	Ac(%)	Se (%)	Es (%)	ROC
(TARTAR <i>et al.</i> , 2013)	RD, CMF, IU	63	90,7	89,6	87,5	0,895
(FARAG <i>et al.</i> , 2011)	ELCAP/LI	70	-	86	97	-
<b>Análise com descritores de forma</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>98,15</b>	<b>96,75</b>	<b>98,87</b>	<b>0,967</b>
<b>Análise com descritores de textura</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>99,57</b>	<b>99,45</b>	<b>99,61</b>	<b>0,992</b>
<b>Todos os descritores</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>99,31</b>	<b>99,59</b>	<b>99,55</b>	<b>0,993</b>
<b>Seleção de características</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>99,09</b>	<b>99,12</b>	<b>99,05</b>	<b>0,990</b>

Comparando o melhor resultado alcançado pelos métodos propostos com os apresentados na Tabela 4.5, o melhor resultado se mostra promissor, uma vez que apresenta valores superiores em todas as métricas de validação em relação a todos os trabalhos de mesma natureza. Além disso, se comparados os números de casos analisados, o método proposto analisou uma quantidade superior de amostras, pois consideramos de extrema importância que os testes sejam realizados com o maior número de casos, garantindo que os métodos propostos sejam os mais genéricos possíveis.

A Tabela 4.6 apresenta os resultados para os métodos que utilizaram medidas de forma e textura para caracterização de seus VOIs. Analisando os resultados, tendo como parâmetro de comparação o método proposto de análise de forma, apenas o trabalho de Lee *et al.* (2010) apresenta um valor um pouco superior na sensibilidade, entretanto, possui valores de acurácia, especificidade e/ou número de exames analisados inferiores ao do estudo proposto.

Quando o comparativo é feito tendo como base o método de análise de textura, os resultados de acurácia, especificidade e sensibilidade obtidos são superiores a todos os outros trabalhos apresentados na Tabela 4.6.

Tabela 4.6: Comparação com outros trabalhos que utilizam descritores de forma e textura para redução de falsos positivos. Legenda: acurácia (Ac), sensibilidade (Se), especificidade (Es) e área sob a curva ROC

<b>Trabalho</b>	<b>Base</b>	<b>Amostra</b>	<b>Ac(%)</b>	<b>Se (%)</b>	<b>Es (%)</b>	<b>ROC</b>
(TERAMOTO <i>et al.</i> , 2016)	-	104	-	90,1	-	-
(ESKANDARIAN; BAGHERZADEH, 2015)	LIDC	147	97	90	-	0,956
(SETIO <i>et al.</i> , 2016)	LIDC	888	-	90,12	-	0,943
(AKRAM <i>et al.</i> , 2015)	LIDC	40	-	96,31	-	-
(SIVAKUMAR; CHANDRASEKAR, 2013)	LIDC-IDRI	54	80,36	82,05	76,47	-
(FILHO <i>et al.</i> , 2013)	LIDC-IDRI	833	97,55	85,91	97,7	0,806
(AL-ABSI <i>et al.</i> , 2012)	JSRT	246	96,58	-	-	-
(OROZCO <i>et al.</i> , 2012)	NBIA/ELCAP	75	82,66	96,15	52,17	-
(ZINOVEV <i>et al.</i> , 2011)	LIDC	32	74,10	-	-	0,690
(LEE <i>et al.</i> , 2010)	LIDC	32	-	97,33	97,11	0,978
(JING <i>et al.</i> , 2010)	3AHG	50	84,39	-	92	-
(LIU <i>et al.</i> , 2009)	NSRCT-LUNG/LIDC	60	87,8	93,25	87,6	-
(HORSTHEMKE <i>et al.</i> , 2009)	JSRT	85	96,28	-	-	-
(HARDIE <i>et al.</i> , 2008)	LIDC	167	70	-	-	-
<b>Análise com descritores de forma</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>98,15</b>	<b>96,75</b>	<b>98,87</b>	<b>0,967</b>
<b>Análise com descritores de textura</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>99,57</b>	<b>99,45</b>	<b>99,61</b>	<b>0,992</b>
<b>Todos os descritores</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>99,31</b>	<b>99,59</b>	<b>99,55</b>	<b>0,993</b>
<b>Seleção de características</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>99,09</b>	<b>99,12</b>	<b>99,05</b>	<b>0,990</b>



Vale salientar que nenhum desses trabalhos utilizaram apenas medidas de forma para caracterização de seus VOIs, assim, não é possível estabelecer uma comparação fidedigna, isto é, apesar de terem a mesma finalidade, utilizam-se de métodos distintos para se alcançar os melhores resultados.

Em resumo, os métodos propostos alcançaram uma acurácia superior ou igual aos melhores resultados encontrados na literatura recente para a redução de falsos positivos de VOIs em TC.

## 4.5 Considerações Finais

Esse capítulo apresentou e detalhou as etapas seguidas pelos métodos propostos, sendo elas divididas em análise de forma e textura. Foram apresentados os resultados para cada método, assim como uma discussão acerca dos valores encontrados. Por fim, fez-se um comparativo com outros trabalhos encontrados na literatura que tratam o problema de redução de falsos positivos.

Vale ressaltar que a delimitação das regiões se mostrou eficiente, melhorando os resultados. Cabe destacar também a eficiência da ME $\alpha$  em selecionar os indivíduos para criação do melhor modelo de treinamento.

No próximo capítulo, serão apresentados os métodos para a classificação de nódulo pulmonar, em maligno e benigno, assim como seus resultados, as discussões e as comparações com os trabalhos relacionados.

# Diagnóstico de nódulos pulmonares

---

Para este capítulo, tem-se a descrição dos procedimentos adotados para o desenvolvimento dos métodos propostos para classificação de nódulo, em maligno e benigno. Em seguida, os resultados, as discussões e, por fim, uma comparação com outros trabalhos da área

## 5.1 Método

Os métodos propostos neste trabalho para classificação de nódulo pulmonar, em maligno e benigno, possuem seu funcionamento separado em 4 etapas. Na Figura 5.1, tem-se um resumo destas etapas. Em suma, na etapa 1) é feita aquisição das imagens; em 2) tem-se a extração e a seleção de características para os nódulos; em 3) realiza-se a seleção do melhor modelo e a classificação e, por fim, 4) é feita a validação e análise dos resultados.

### 5.1.1 Aquisição de Imagens

Como mencionado na Seção 4.1.1, a base de imagens utilizada para validação dos métodos é a LIDC-IDRI (ARMATO *et al.*, 2011). Essa base possui 1.018 exames de TC. Sendo que, apenas 833 exames foram utilizados para aplicação do método.

Para compor a base de imagens para testes dos métodos de diagnóstico, teve-se como princípio o trabalho desenvolvido por Jabon *et al.* (2009). Isto se

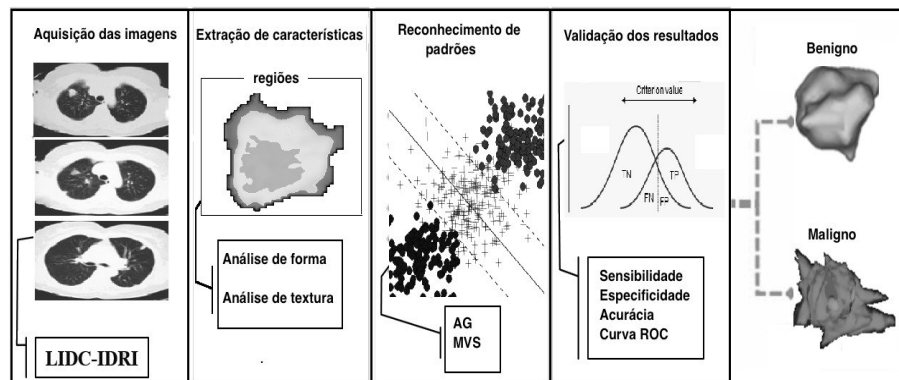


Figura 5.1: Esquema geral seguido pelos métodos propostos, para diagnóstico de nódulo pulmonar.

deu pelo fato que não há imposição para que haja consenso, já que todos os nódulos indicados pela revisão dos radiologistas são averiguados e salvos. Com isso, torna-se possível que um mesmo nódulo possua diferentes diagnósticos.

Baseado nisso, considera-se então, para compor esta base, 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 contornar esse problema, 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 5.2 ilustra o processo:

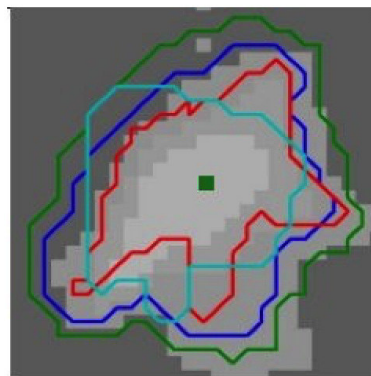
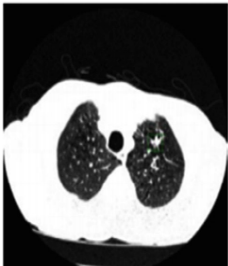


Figura 5.2: Ilustração do resumo dos nódulos. Fonte (JABON *et al.*, 2009).

Na Figura 5.2, as linhas coloridas representam as marcações definidas individualmente pelos especialistas. O quadrado verde no centro refere-se ao

centróide calculado para o contorno de mesma cor. Caso as coordenadas desse centróide se encontrem nas áreas delimitadas por outros especialistas, considera-se neste trabalho que se trata do mesmo nódulo e, portanto, só deve ser aceita uma instância que será referente àquela marcação do nódulo que possuir maior área de contorno. No exemplo simplificado da Figura 5.2, a instância do nódulo que será utilizada será aquela representada pelo contorno de cor verde, pois atende a esse critério.

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 no que diz respeito à malignidade ou benignidade. A avaliação do diagnóstico já está presente para cada nódulo salvo na base LIDC-IDRI, em uma escala subjetiva de cinco níveis que varia de altamente suspeito de benignidade a altamente suspeito de malignidade, isto é, quão mais próximo de 1, mais tendencioso a ser benigno e quanto mais próximo de 5, maior a probabilidade de maligno. Para resumir essas informações, utilizam-se as informações conforme levantamento da etapa anterior para que seja efetuado cálculo segundo apresentado em (JABON *et al.*, 2009), onde os valores das características pertencentes ao mesmo nódulo são reduzidos a um único valor através do cálculo da moda ou da mediana. A mediana só é usada para resumir o valor do grau de malignidade quando a moda se repete. Esse processo é ilustrado na Figura 5.3, destacando com o retângulo a característica de malignidade que será utilizada na composição da base de imagens.



Rad.	Lob.	Mal.	Marg.	Spher.	Spic.	Subt.	Text.
A	3	4	4	2	4	3	4
B	4	3	4	4	3	5	5
C	4	2	3	4	3	4	5
D	4	3	2	2	4	3	3
Summarized	4	3	4	3	3	3	5

Figura 5.3: Exemplo do resumo do diagnóstico dos nódulos. Fonte (JABON *et al.*, 2009).

Para não tornar o trabalho repetitivo, cabe esclarecer que todas as demais etapas do método - Definição das regiões internas, Extração de características,

Reconhecimento de padrões e Validação dos resultados - foram as mesmas seguidas pelos métodos de redução de falsos positivos mostrados no Capítulo 4. Entretanto, algumas definições serão trocadas por termos comumente utilizados em trabalhos que tratam o problema de diagnóstico pulmonar, como: VOI será nódulo pulmonar, e as classes de nódulo e não-nódulo passarão a ser, maligno e benigno, respectivamente.

## 5.2 Resultados

Nesta seção, serão apresentados os resultados dos descritores propostos para classificação de nódulos, em maligno e benigno. Para tanto, dos 833 exames disponíveis para validação dos métodos, foram extraídos 1405 nódulos (1011 benignos e 394 malignos) (Seção 5.1.1).

O critério de teste e treino seguiu o mesmo apresentado na Seção 4.2, que trata de redução de falsos positivos. Em resumo, para os experimentos com objetivo de classificar o nódulo, em maligno e benigno, dividiu-se em 80% (1124) para seleção do melhor modelo e 20% (281) para teste, todos selecionados de forma aleatória. Esses 20% serão usados para testar o modelo final selecionado pela ME $\alpha$ .

### 5.2.1 Resultados para os Descritores de Textura

O ideal para um sistema CAD ou CADx é ter um bom equilíbrio entre as três métricas de avaliação (acurácia, sensibilidade e especificidade), pois, um bom método deve ser capaz de classificar com sucesso, tantos os casos de positivos (que possuem a doença) como os casos de negativos (que não possuem a doença). Nas Tabelas 5.1 e 5.2, são apresentados os resultados para a classificação entre maligno e benigno dos nódulos pulmonares, usando apenas descritores de textura.

Nos resultados apresentados na Tabela 5.1, foi analisado apenas o nódulo pulmonar.

Analisando os resultados da Tabela 5.1, tendo como base o critério do equilíbrio entre as métricas de desempenho, o melhor resultado está quando combinados todos os índices, alcançando acurácia de 88,61%, sensibilidade de 93,10%, especificidade de 87,02% e uma área sob a curva ROC de 0,809. No pior

Tabela 5.1: Resultados para o diagnóstico sem o uso das regiões, usando análise de textura, baseado em índices de diversidade filogenética.

Índices baseados em:	Acurácia	Sensibilidade	Especificidade	ROC	Número de características
Riqueza de espécies	83,98	93,10	79,89	0,867	2
Distância entre pares de espécies	88,61	86,20	89,69	0,837	5
Topologia	87,90	89,65	87,11	0,881	2
Caminho mínimo	83,27	91,95	79,38	0,856	3
<b>Todos os índices</b>	<b>88,61</b>	<b>93,10</b>	<b>87,02</b>	<b>0,898</b>	<b>12</b>

caso, pode-se destacar os resultados dos índices baseados no caminho mínimo, que alcançaram uma acurácia de 83,27%, uma sensibilidade de 91,95%, uma especificidade de 79,38% e, por fim, uma curva ROC de 0,856.

A Tabela 5.2 apresenta os resultados para análise dos índices de diversidade, usando o nódulo pulmonar e as 3 regiões geradas a partir do mesmo.

Tabela 5.2: Resultados para o diagnóstico com o uso das regiões, usando análise de textura, baseado em índices de diversidade filogenética.

Índices baseados em:	Acurácia	Sensibilidade	Especificidade	ROC	Número de características
Riqueza de espécies	86,12	86,07	86,13	0,864	8
Distância entre pares de espécies	91,40	91,98	91,14	0,921	20
Topologia	90,67	91,50	90,34	0,902	8
Caminho mínimo	88,23	87,75	87,90	0,878	12
<b>Todos os índices</b>	<b>92,52</b>	<b>93,10</b>	<b>92,26</b>	<b>0,921</b>	<b>48</b>

A Tabela 5.2 apresenta o quão promissor são os métodos propostos por esta tese. No melhor caso, tem-se quando todos os índices são combinados para descrever as regiões. Com isso, alcançou-se uma acurácia de 92,52%, sensibilidade de 93,1%, especificidade de 92,26% e uma área sob curva ROC de 0,921. No pior caso, destacam-se os resultados do grupo de índice baseado na riqueza de espécies.

Acredita-se que o melhor resultado, deu-se porque os índices em conjunto tornaram mais evidentes propriedades que até então um grupo de índice isoladamente não seria capaz de evidenciar. Já no pior caso, o fato responsável por um pior desempenho está relacionado a análise apenas da riqueza de espécies, isso porque ela não fornece propriedades para descrever de maneira precisa a textura dos nódulos. Em outras palavras, acredita-se que as propriedades dos

índices que mensuram a cerca da semelhança entre as espécies, isto é, relações filogenéticas/grau de parentesco, conseguem descrever com uma maior riqueza de detalhes uma determinada região analisada.

## 5.2.2 Resultados para os Descritores de Forma

Nesta seção, são tabulados os resultados usando descritores de forma para diferenciação dos padrões de malignidade e benignidade.

Inicialmente, seguindo os passos de demonstração dos resultados, são realizados testes que analisaram apenas o nódulo.

Tabela 5.3: Resultados dos descritores de forma para o diagnóstico, analisando o nódulo sem o uso da região intralesional.

Descritor	Acurácia	Sensibilidade	Especificidade	ROC	Número de características
Diagramas de forma	82,77	78,64	84,40	0,778	26
Abordagem cilíndrica	85,3	80,59	86,79	0,809	6
Funcionais de Minkowski	83,87	79,74	85,50	0,804	4
Medidas de distância	87,03	86,90	87,69	0,839	2
Medidas do esqueleto	86,02	82,55	87,56	0,837	6
Medidas de triangulação	84,94	81,81	86,13	0,804	2
Medidas baseadas nos diâmetros de Feret	86,73	89,20	85,78	0,874	4
<b>Todos os descritores</b>	<b>89,24</b>	<b>87,50</b>	<b>89,94</b>	<b>0,879</b>	<b>50</b>

A Tabela 5.3 apresenta o resultado para o diagnóstico, usando somente descritores de forma. No melhor caso, mais uma vez, está quando combinadas todas as técnicas. Em suma, obteve-se uma acurácia de 89,24, sensibilidade e especificidade de 87,5 e 89,94, respectivamente, e uma área sob a curva ROC de 0,879. No pior caso, destacam-se as medidas baseadas nos diagramas de forma, com uma acurácia de 82,77% e uma curva ROC de 0,778.

Nos testes seguintes, foram utilizadas para extração das características o nódulo e a região intralesional (subregiões).

Tabela 5.4: Resultados dos descritores de forma para o diagnóstico analisando o nódulo e a região intralesional.

Descritor	Acurácia	Sensibilidade	Especificidade	ROC	Número de características
Diagramas de forma	84,10	83,50	84,36	0,831	52
Abordagem cilíndrica	86,02	82,55	87,56	0,821	12
Funcionais de Minkowski	85,30	84,70	85,56	0,848	8
Medidas de distância	91,03	90,54	91,21	0,913	4
Medidas do esqueleto	85,30	80,59	86,79	0,829	12
Medidas de triangulação	88,53	85,91	89,42	0,874	4
Medidas baseadas nos diâmetros de Feret	90,68	87,84	92,11	0,896	8
<b>Todos os descritores</b>	<b>93,19</b>	<b>92,75</b>	<b>93,33</b>	<b>0,930</b>	<b>100</b>

Os resultados apresentados na Tabela 5.4 demonstram a efetividade dos descritores de forma propostos. Novamente, destacam-se os resultados quando combinados todos os descritores, alcançando valores expressivos e promissores. Alcançando uma acurácia, sensibilidade, especificidade e curva ROC de 93,19%, 92,75%, 93,33% e 0,930, respectivamente.

Como pode ser observado na Tabela 5.4, os resultados com a análise da região intralesional proporcionaram uma melhora nos resultados, ou seja, a região permitiu extrair informações da região interna ao nódulo capaz de melhorar o desempenho dos descritores, com isso, quando se combinou todos os descritores, os resultados apresentaram uma melhora significativa.

Analisando os resultados apresentados nas Tabelas 5.3 e 5.4, é possível concluir que a quantidade de característica não possui uma relação direta com os resultados, uma vez que, os resultados do grupo de descritores baseados em distância são capazes de alcançar valores bem superiores com apenas 4 características, se comparadas, por exemplo, aos resultados dos diagramas de forma, que possuem 52 medidas.



### 5.2.3 Resultados com todos os descritores e seleção de características

Assim como na tarefa de redução de falsos positivos, no diagnóstico não é diferente, pois as análises de forma e textura teoricamente se complementam. Com base nisso, foi realizado um teste para verificar a eficiência dos descritores em conjunto. Nesse teste, cada nódulo é representado por 148 características. Os resultados apresentaram uma pequena melhora, atingindo uma acurácia de 93,46%, sensibilidade de 92,95%, especificidade de 93,49% e uma área sob a curva ROC de 0,931.

Para averiguar o potencial dos descritores, é feita uma seleção de características usando a Análise Discriminante *Stepwise*.

Após a seleção, resultaram apenas 14 características, sendo elas: 4 dos índices baseados na distância entre pares de espécies; 3 dos índices baseados na topologia; 4 das medidas de distância; 1 da abordagem cilíndrica; 1 das medidas baseadas na triangulação; 1 dos diâmetros de Feret. Acredita-se que estas medidas foram selecionadas pelo método *Stepwise* devido a sua capacidade de caracterização, ou seja, conseguem fornecer mais detalhes do objeto analisado, como: o grau de parentesco que as espécies possuem e averiguar o quão regular esfericamente o objeto possa ser.

Os resultados foram satisfatórios, pois apenas 14 medidas foram usadas para caracterizar o nódulo e os resultados atingidos foram bem próximos ao melhor caso. Assim, alcançou-se uma acurácia de 92,24%, uma sensibilidade de 92,01%, uma especificidade de 92,45% e uma curva ROC de 0,919. Como pode ser observado, o resultado é um pouco abaixo do melhor caso, entretanto, é promissor, pois foi possível alcançar uma taxa de acerto acima dos 92% com um número de descritores bem reduzido.

Por fim, na Figura 5.4, são apresentadas as curvas ROC para os melhores resultados dos métodos de diagnóstico pulmonar para as análises de: textura, forma, todas as medidas em conjunto e após a seleção de características.

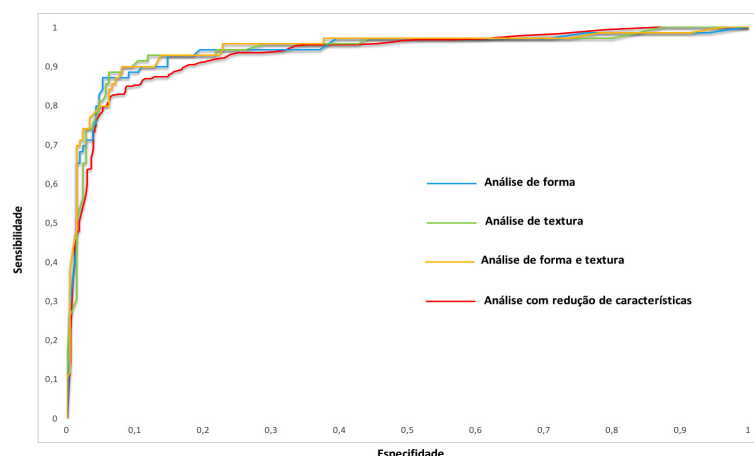


Figura 5.4: Resultados para as curvas ROC dos métodos de diagnóstico pulmonar.

### 5.3 Discussão

Partindo da premissa que um sistema CAD ou CADx deve apresentar bons resultados nas métricas que aferem seu desempenho - como: acurácia, sensibilidade e especificidade -, isto é, considerando o padrão *gold*, estimado em média de 85% (BRAZ JUNIOR *et al.*, 2013), a busca por métodos que elevam as taxas de acerto é cada vez mais presente.

Seguindo com a análise dos resultados, empregaram-se os métodos propostos para resolução de outro problema característico de sistemas CADx. Para isso, aplicaram-se os métodos com o intuito de classificar os nódulos, no que diz respeito às classes maligno e benigno. Com isso, têm-se os resultados das Tabelas 5.1, 5.2, 5.3 e 5.4.

Primeiramente, realizaram-se os experimentos usando apenas os descritores de textura e analisando apenas o nódulo. Partindo do valor estabelecido para o padrão *gold*, o melhor resultado, dos apresentados na Tabela 5.1, está quando combinam todos os índices, pois foram alcançados valores aceitáveis, mas um pouco longe do desejado. Com isso, foram feitos novos experimentos e apresentados na Tabela 5.2. Esses experimentos foram realizados usando além do nódulo, as regiões geradas a partir dele. Essa abordagem das regiões proporcionou uma melhora significativa nos resultados, em especial, quando combinados todos os índices, alcançando valores expressivos acima dos 92% em todas as métricas de avaliação do desempenho. Acredita-se que a análise

da região intralesional, isto é, regiões internas, tenham contribuído fortemente para tal.

E por último, têm-se os experimentos apresentados nas Tabelas 5.3 e 5.4. Nesses experimentos, foi possível averiguar as propriedades geométricas não apenas do nódulo em si, mas também de sua região intralesional. Na Tabela 5.3, os descritores de forma foram aplicados apenas sobre o nódulo. Tomando por base o valor do padrão *gold*, os resultados são promissores, pois em todas as métricas de avaliação os resultados estão acima dos 87%. É importante destacar que dentre todas as medidas de forma, as medidas que alcançaram um melhor resultado foram as baseadas na distância, pois, mostram-se mais sensíveis as alterações encontradas no nódulo.

Por conta de possíveis excessos devido a divergências das marcações intra-especialistas, os nódulos, sejam eles benignos ou malignos, possuem formas bem próximas. Partindo deste, têm-se os resultados dos experimentos usando além do nódulo, a região intralesional (Tabela 5.4). Com isso, os resultados alcançados, mais uma vez foram melhorados, atingindo valores nas métricas de avaliação acima dos 92%.

É importante destacar a melhora nos resultados quando é usado a região interna ao nódulo, pois, foi possível analisar o comportamento da forma e da textura da região interna de maneira isolada, tornando possível averiguar as mudanças ou variações das células dos nódulos malignos e benignos de modo a estabelecer um padrão de diferença entre as classes.

Vale salientar que a base LIDC-IDRI não dispõe de exames histopatológicos para comprovação da malignidade do nódulo pulmonar, este por sua vez, tem sua classificação baseada apenas em informações dos quatro radiologistas. Além disso, a base LIDC-IDRI não dispões de exames em séries temporais para um acompanhamento de uma suspeita de malignidade. Estas, são limitações encontradas na base de imagens usada.

Os métodos propostos demonstram-se muito promissores, uma vez que, na diferenciação de malignidade e benignidade, os resultados foram bem acima do padrão *gold*, superando os 92% de acerto total.

Cabe destacar que, quando se combinam os descritores de textura e forma, os resultados são ligeiramente melhorados, entretanto, com um número

um pouco elevado de características para o tamanho da amostra estudada. Uma vez que, há correlação entre os descritores. Com isso, após a seleção das melhores características, os resultados continuaram consistentes, pois, permaneceram com valores acima dos 92% e com um número de descritores bem reduzidos.

Esses bons resultados demonstram também o quão é importante a análise interna ao nódulo. Em todos os casos, proporcionou bons resultados, conseguindo melhorar os resultados de forma considerável. Além desses, os bons resultados asseguram a efetividade da ME $\alpha$  em selecionar os melhores indivíduos para composição do melhor modelo.

## 5.4 Comparação com outros Trabalhos

Será apresentado a seguir um comparativo com outros trabalhos encontrados na literatura que tratam sobre o mesmo problema dos métodos propostos para diagnóstico de nódulo pulmonar. A comparação com outros trabalhos não é uma tarefa fácil, isto porque nenhum dos trabalhos relatados nesta tese forneceu os exames utilizados. Poucas informações como, banco de imagens utilizado, são disponibilizadas, e isto nem é encontrado para todos os trabalhos, além disso, existem bases privadas que não são disponibilizadas. Assim, tornou-se inviável realizar uma avaliação rigorosa de nossos métodos em relação a outras obras.

Nas Tabelas 4.5 e 4.6, tem-se uma visão resumida (base de exames usado, métricas de avaliação de desempenho) dos resultados encontrados nos trabalhos relacionados. Assim, pretende-se mostrar que os métodos propostos são promissores, uma vez que, em comparação com outras obras, alcançaram resultados acima do valor do padrão *gold*, em torno de 92% e 93% para a tarefa de diagnóstico de nódulo pulmonar em maligno e benigno.

Seguindo com a comparação dos métodos propostos, comparam-se agora os resultados para as metodologias CADx, que possuem a finalidade de classificar o nódulo em maligno ou benigno. Primeiramente, comparam-se os resultados com outros métodos que utilizam descritores de textura e forma para caracterização de seus nódulos (Tabela 5.5). E por fim, tem-se a comparação com métodos que utilizam apenas descritores de forma (Tabela 5.6).

Comparando os resultados alcançados dos métodos propostos, com os

Tabela 5.5: Comparação com outros trabalhos que utilizam apenas descritores de textura para diagnóstico. Legenda: acurácia (Ac), sensibilidade (Se), especificidade (Es) e área sob a curva ROC

Trabalho	Base	Amostra	Ac(%)	Se (%)	Es (%)	ROC
(FIRMINO <i>et al.</i> , 2016)	LIDC-IDRI	1109	-	93,1	88,14	0,91
(KUMAR <i>et al.</i> , 2015)	ELCAP e NBIA	128	84	-	-	-
(OROZCO <i>et al.</i> , 2015)	ELCAP e LIDC	106	82	90,9	73,91	0,805
(DANDIL <i>et al.</i> , 2014)	Privado	128	90,63	92,30	89,47	-
(PARVEEN; KAVITHA, 2014)	Privado	11	-	91,38	89,56	-
(KREWER <i>et al.</i> , 2013)	LIDC-IDRI	33	90,91	85,71	94,74	-
(ZHANG <i>et al.</i> , 2013)	ELCAP	78	-	-	-	-
(OROZCO <i>et al.</i> , 2013)	LIDC e NCI	4323	75,01	83,35	-	-
(NASCIMENTO <i>et al.</i> , 2012)	LIDC-IDRI	73	83,75	82,95	84,58	-
(OROZCO <i>et al.</i> , 2012)	NBIA-ELCAP	-	82,66	96,15	52,17	-
<b>Análise com descritores de forma</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>93,19</b>	<b>92,75</b>	<b>93,33</b>	<b>0,930</b>
<b>Análise com descritores de textura</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>92,52</b>	<b>93,1</b>	<b>92,26</b>	<b>0,921</b>
<b>Todos os descritores</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>93,46</b>	<b>92,95</b>	<b>93,49</b>	<b>0,931</b>
<b>Seleção de características</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>92,24</b>	<b>92,01</b>	<b>92,45</b>	<b>0,919</b>

apresentados na Tabela 5.5, os resultados alcançados mostram-se promissores. Quando se analisado o equilíbrio entre as três métricas de desempenho, as soluções para diagnóstico pulmonar propostas mostram-se mais robustas, pois conseguem resultados acima dos 92% para os valores de acurácia, sensibilidade e especificidade. Além disso, apenas o trabalho de Orozco *et al.* (2013) apresenta um número superior de casos analisados, entretanto, seus resultados são bem inferiores aos alcançados.

Tabela 5.6: Comparação com outros trabalhos que utilizam descritores de forma e textura para diagnóstico. Legenda: acurácia (Ac), sensibilidade (Se), especificidade (Es) e área sob a curva ROC

Trabalho	Base	Amostra	Ac(%)	Se (%)	Es (%)	ROC
(DHARA <i>et al.</i> , 2016)	LIDC-IDRI	891	-	89,73	74,9	0,882
(ELIZABETH <i>et al.</i> , 2012)	Privado	-	94,44	-	88,14	-
(KUMAR <i>et al.</i> , 2011)	Privado	-	90	86	84	-
(EL-BAZ <i>et al.</i> , 2011)	Privado	327	93,6	-	-	-
(WAY <i>et al.</i> , 2009)	Privado	256	85	-	-	0,857
<b>Análise com descritores de forma</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>93,19</b>	<b>92,75</b>	<b>93,33</b>	<b>0,930</b>
<b>Análise com descritores de textura</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>92,52</b>	<b>93,1</b>	<b>92,26</b>	<b>0,921</b>
<b>Todos os descritores</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>93,46</b>	<b>92,95</b>	<b>93,49</b>	<b>0,931</b>
<b>Seleção de características</b>	<b>LIDC-IDRI</b>	<b>833</b>	<b>92,24</b>	<b>92,01</b>	<b>92,45</b>	<b>0,919</b>

A Tabela 5.6 apresenta os resultados de trabalhos que utilizam medidas de forma e textura para caracterização do nódulo pulmonar. Os trabalhos de El-Baz *et al.* (2011) e Elizabeth *et al.* (2012) apresentam uma acurácia um pouco superior às dos métodos propostos, porém isso não implica afirmar que sejam mais eficiente, pois, valores de sensibilidade e especificidade que medem o quão eficiente é um método em termos de diagnosticar ou não a presença da doença são bem inferiores ou não informados. Além disso, o número de casos analisados pelos autores é bem inferior.

## 5.5 Considerações Finais

Nesse capítulo, foram apresentados e discutidos os resultados produzidos pelos experimentos realizados utilizando os métodos propostos para diagnóstico de nódulos em maligno e benigno. Também realizou-se uma comparação dos resultados obtidos com outros presentes na literatura.

Assim como na redução de falsos positivos, a estratégia de delimitar regiões se mostrou eficaz, pois em todos os experimentos, melhoraram os resultados alcançados. Cabe destacar também a eficiência da  $ME\alpha$  em selecionar os melhores indivíduos para compor o melhor modelo de treinamento.

No Capítulo 6, serão feitas algumas conclusões acerca da pesquisa desenvolvida nesta tese, bem como apresentadas sugestões de trabalhos futuros.

## CAPÍTULO 6

# Conclusão

---

A redução de falsos positivos, assim como o diagnóstico, são tarefas de complexidade elevada e de grande relevância, dado pelo fato de que os VOIs de nódulo pulmonar, muitas vezes possuem forma e textura muito semelhantes, mesmo sendo de classes distintas. Além disso, o nódulo pulmonar devido aos seus tipos podem possuir comportamentos que, em alguns casos, podem provocar erros na classificação, como por exemplo os casos dos hamartomas, que são nódulos benignos, mas que possuem calcificações com padrões em formato de pipoca e localização complexa.

Os resultados foram mais promissores para tarefa de redução de falsos positivos devido a complexidade do problema, uma vez que, se comparado com a tarefa de classificação de nódulo em maligno e benigno, a redução de falsos positivos é menos complexa, devido as propriedades presentes nos nódulos serem bem definidas, se comparadas aos não-nódulos. É importante destacar que até para o próprio especialista médico, a tarefa de classificação em maligno e benigno apresenta uma complexidade tão elevada que é necessário acompanhamento e biópsias durante um determinado tempo para de fato estabelecer um diagnóstico para a lesão.

Em suma, esta tese apresentou métodos para a redução de falsos positivos e o diagnóstico de nódulo pulmonar em tomografia computadorizada através de técnicas de processamento de imagens e reconhecimento de padrões, explorando as propriedades de forma e textura. Nesse âmbito, foi apresentada a fundamentação teórica dos assuntos relacionados a área de estudo, os quais são necessários para



a compreensão do trabalho desenvolvido e também das técnicas abordadas nos métodos. Também são apresentados alguns trabalhos relacionados nas áreas de redução de falsos positivos e diagnóstico.

Os métodos propostos foram validados através de experimentos usando a base LIDC-IDRI. Em resumo, alguns descritores merecem destaque nos métodos propostos, consideradas também como as principais contribuições desta tese. A primeira é a adaptação e o uso dos conceitos da biologia, mais precisamente dos índices de diversidade filogenética e das árvores filogenéticas; para uso no âmbito computacional, especificamente no processamento de imagens para descrição de textura. A segunda foi o desenvolvimento e a adaptação de descritores que aferem informações da forma da região analisada. E por fim, pode-se destacar a exploração de uma forma automática de delimitar regiões internas ao VOI, possibilitando novas formas de análise.

A base LIDC-IDRI é uma base extremamente complexa e diversificada. Ela contém inúmeros casos de nódulos pulmonares. Partindo do princípio de sua formação realizada pelo consórcio entre o *Lung Image Database Consortium e a Image Database Resource Initiative*, essa base dispõe de exames que foram extraídos por diversos tomógrafos, dificultando os processos dos sistemas CAD e CADx.

Os resultados produzidos pelos índices de diversidade filogenética em conjunto com as árvores filogenéticas demonstram resultados muito promissores, tanto na redução de falsos positivos quando no diagnóstico. Além desses, cabe ressaltar a eficiência dos descritores de forma que, com uma abordagem diferente de delimitação da região analisada, mostraram-se capazes de alcançar resultados tão expressivos quanto os descritores de textura, seja na redução de falsos positivos ou no diagnóstico.

Finalmente, esta tese apresenta como contribuições os seguintes itens:

- O desenvolvimento de métodos para redução de falsos positivos e diagnóstico usando análises de forma e textura;
- Utilização de índices usados em outras áreas do conhecimento para caracterização da textura dos volumes extraídos das TCs para redução de falsos positivos e diagnóstico de nódulo pulmonar:

- Utilização dos índices baseados na riqueza de espécies;
  - Utilização dos índices baseados na distância entre pares de espécies;
  - Utilização dos índices baseados na topologia; e
  - Utilização dos índices baseados no caminho mínimo.
- Construção de uma representação da imagem através de árvores filogenéticas para uso dos índices de diversidade filogenéticos;
  - Na adaptação e no desenvolvimento de descritores para extrair informações da forma dos volumes extraídos das TCs para redução de falsos positivos e diagnóstico de nódulo pulmonar:
    - Adaptação dos diagramas de forma e das funcionais de Minkowski;
    - Desenvolvimento de:
      - \* Medidas de proporção;
      - \* Abordagem baseada em cilindros;
      - \* Descritores baseados em distância;
      - \* Descritores baseados no esqueleto;
      - \* Descritores baseados em triangulação; e
      - \* Descritores baseados nos diâmetros de Feret.
  - Delimitação automática de regiões internas ao volume de interesse que possibilitou a extração de informações da textura e da forma da região intralesional; e
  - Desenvolvimento de diretrizes para compor um sistema CAD na etapa de redução de falsos positivos e para diagnósticos do câncer de pulmão a partir de nódulos em TC, provendo ao especialista uma segunda opinião nas análises.

Em linhas gerais, os métodos descritos nesta tese poderão compor uma ferramenta CAD e CADx a ser aplicada em casos reais de detecção e acompanhamento do desenvolvimento de nódulos pulmonares a partir de imagens de TC. Vale ainda analisar a robustez das mesmas ao serem testadas em uma base de imagens com um grande número de casos e com dificuldades de análise.

## 6.1 Trabalhos Futuros

Como exposto nesta tese, o interesse em pesquisas de cunho computacional que auxiliem na área médica é de extrema relevância, dada a grande quantidade de pesquisas existentes. Logo, como trabalhos futuros, enumeram-se algumas sugestões:

1. Construção das ferramentas CADx integrando os métodos desenvolvidos por esta tese;
2. Integração dos métodos de redução de falsos positivos a sistemas CADs;
3. Estudo e análise de novos índices de diversidade, como os índices funcionais, que possuem como objetivo analisar não apenas as relações entre as espécies, mas também possíveis características, como evolução e outras;
4. Utilizar mais recursos de processamento de imagens, como técnicas de melhoramento de contraste;
5. Utilizar uma nova representação da imagem, como *Local Binary Pattern*;
6. Desenvolver novas formas de representação da árvores filogenética, como por exemplo, o uso do cladograma;
7. Testar os métodos propostos em outras bases de imagens, com o intuito de verificar a sua robustez;
8. Testar os métodos em VOIs identificados como nódulos não-sólidos;
9. Aplicar os descritores de textura e forma para análises de lâminas de biópsias de nódulo pulmonar;
10. Aplicar os métodos propostos para a caracterização de textura e forma de anormalidades de outros tipos de imagens, por exemplo, de mama; e
11. Aplicar os métodos propostos em imagens de Ressonância Magnética.

# Referências

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>.

AL-ABSI, H.; SAMIR, B.; SHABAN, K.; SULAIMAN, S. Computer aided diagnosis system based on machine learning techniques for lung cancer. In: *International Conference on Computer Information Science (ICIS)*. [S.l.: s.n.], 2012. v. 1, p. 295–300.

ALAJLAN, N.; KAMEL, M. S.; FREEMAN, G. Multi-object image retrieval based on shape and topology. *Image Commun.*, Elsevier Science Inc., New York, NY, USA, v. 21, n. 10, p. 904–918, nov. 2006. ISSN 0923-5965. Disponível em: <http://dx.doi.org/10.1016/j.image.2006.09.002>.

ALAJLAN, N.; RUBE, I. E.; KAMEL, M. S.; FREEMAN, G. Shape retrieval using triangle-area representation and dynamic space warping. *Pattern Recognition*, v. 40, n. 7, p. 1911 – 1920, 2007. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320306005188>.

ARMATO, S. G.; MCLENNAN, G.; BIDAUT, L.; MCNITT-GRAY, M. F.; MEYER, C. R.; REEVES, A. P.; ZHAO, B.; ABERLE, D. R.; HENSCHKE, C. I.; HOFFMAN, E. A. *et al.* The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, American Association of Physicists in Medicine, v. 38, n. 2, p. 915–931, 2011.

AZEVEDO, E.; CONCI, A. *Computação gráfica: teoria e prática*. CAMPUS - RJ, 2003. ISBN 9788535212525. Disponível em: <https://books.google.com.br/books?id=Y1mhBAAACAAJ>.

BAXEVANIS, A. D.; OUELLETTE, B. F. F. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. Wiley, 2004. (Methods of Biochemical Analysis). ISBN 9780471461012. Disponível em: [http://books.google.com.br/books?id=ghU\\\_-ZaEAdHUC](http://books.google.com.br/books?id=ghU\_-ZaEAdHUC).

BISHOP, C. *Pattern recognition and machine learning*. Secaucus, NJ: Springer-Verlag, 2006. ISBN 9780387310732. Disponível em: <http://www.springer.com/us/book/9780387310732>.

BLUM, H. A Transformation for Extracting New Descriptors of Shape. *Models for the Perception of Speech and Visual Form*, MIT Press, Cambridge, MA, p. 362–380, 1967. Disponível em: <http://pageperso.lif.univ-mrs.fr/~edouard.thiel/rech/1967-blum.pdf>.

BLUMER, A.; EHRENFEUCHT, A.; HAUSSLER, D.; WARMUTH, M. K. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, ACM, New York, NY, USA, v. 36, n. 4, p. 929–965, out. 1989. ISSN 0004-5411. Disponível em: <http://doi.acm.org/10.1145/76359.76371>.

BRAY, T.; PAOLI, J.; SPERBERG-MCQUEEN, C. M.; MALER, E.; YERGEAU, F. Extensible markup language (xml). *World Wide Web Consortium Recommendation REC-xml-19980210*. <http://www.w3.org/TR/1998/REC-xml-19980210>, v. 16, 1998.

BRAZ JUNIOR, G.; ROCHA, S. V. da; GATTASS, M.; SILVA, A. C.; PAIVA, A. C. d. A mass classification using spatial diversity approaches in mammography images for false positive reduction. *Expert Systems with Applications*, Pergamon, v. 40, n. 18, p. 7534–7543, 2013.

BROWN, C. D.; DAVIS, H. T. Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems*, v. 80, n. 1, p. 24 – 38, 2006. ISSN 0169-7439. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0169743905000766>.

CHANG, C.-C.; LIN, C.-J. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, ACM, New York, NY, USA, v. 2, n. 3, p. 27:1–27:27, maio 2011. ISSN 2157-6904. Disponível em: <http://doi.acm.org/10.1145/1961189.1961199>.

CHEN, W.; LI, Z.; BAI, L.; LIN, Y. Nf-kappab in lung cancer, a carcinogenesis mediator and a prevention and therapy target. *Frontiers in bioscience (Landmark edition)*, v. 16, p. 1172–85, 2011. ISSN 1093-4715.

CHEN, W.; LI, Z.; BAI, L.; LIN, Y. Nf-kappab, a mediator for lung carcinogenesis and a target for lung cancer prevention and therapy. *Frontiers in bioscience: a journal and virtual library*, v. 16, p. 1172–1185, 2012.

CLARKE, K. R.; WARWICK, R. M. A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology*, British Ecological Society, v. 35, n. 4, p. pp. 523–531, 1998. ISSN 00218901. Disponível em: <http://www.jstor.org/stable/2405167>.

CLARKE, K. R.; WARWICK, R. M. R. M.; LABORATORY, P. M. Book; Book/Illustrated. *Change in marine communities: an approach to statistical analysis and interpretation*. 2nd ed. ed. [S.l.]: Plymouth, U.K. : PRIMER-E Ltd, 2001. Includes bibliographical references (p. A3-1-A3-5).

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

DAVID, L.; KIEN, K.; MARIE, F. D. Computation of minkowski measures on 2d and 3d binary images. *Image Analysis and Stereology*, v. 26, n. 2, p. 83–92, 2011. ISSN 1854-5165. Disponível em: <http://www.ias-iss.org/ojs/IAS/article/view/811>.

DEHMESHKI, J.; YE, X.; CASIQUE, M. V.; LIN, X. A hybrid approach for automated detection of lung nodules in ct images. In: *ISBI. IEEE*, 2006. p. 506–509. Disponível em: <http://dblp.uni-trier.de/db/conf/isbi/isbi2006.html>.

DEWAN, N. A.; SHEHAN, C. J.; REEB, S. D.; GOBAR, L. S.; SCOTT, W. J.; RYSCHON, K. Likelihood of malignancy in a solitary pulmonary nodule: Comparison of bayesian analysis and results of fdg-pet scan. *Chest*, v. 112, n. 2, p. 416 – 422, 1997. ISSN 0012-3692. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0012369215530109>.

DHARA, A. K.; MUKHOPADHYAY, S.; DUTTA, A.; GARG, M.; KHANDELWAL, N. A combination of shape and texture features for classification of pulmonary nodules in lung ct images. *Journal of Digital Imaging*, p. 1–10, 2016. ISSN 1618-727X. Disponível em: <http://dx.doi.org/10.1007/s10278-015-9857-6>.

DUDA, R. O.; HART, P. E. *Pattern Classification and Scene Analysis*. New York: Wiley-Interscience Publication, 1973. ISBN 0471223611. Disponível em: <http://www.ica.luz.ve/~enava/redesn/ebooks/DHS/Versi%F3n%20PS/DHSChap5.ps>.

EL-BAZ, A.; NITZKEN, M.; KHALIFA, F.; ELNAKIB, A.; GIMELFARB, G.; FALK, R.; EL-GHAR, M. 3d shape analysis for early diagnosis of malignant lung nodules. In: SZEKELY, G.; HAHN, H. (Ed.). *Information Processing in Medical Imaging*. Springer Berlin Heidelberg, 2011. (Lecture Notes in Computer Science, v. 6801), p. 772–783. ISBN 978-3-642-22091-3. Disponível em: <http://dx.doi.org/10.1007/978-3-642-22092-0-63>.

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. *Image Processing, IET*, v. 6, n. 6, p. 697–705, August 2012. ISSN 1751-9659.

- ERASMUS, J. J.; CONNOLLY, J. E.; MCADAMS, H. P.; ROGGLI, V. L. Solitary pulmonary nodules: Part i. morphologic evaluation for differentiation of benign and malignant lesions. *RadioGraphics*, v. 20, n. 1, p. 43–58, 2000. PMID: 10682770. Disponível em: <http://dx.doi.org/10.1148/radiographics.20.1.g00ja0343>.
- ERKEL, A. van; PATTYNAMA, P. Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. *European Journal of Radiology*, v. 27, n. 2, p. 88–94, 1998.
- ESKANDARIAN, P.; BAGHERZADEH, J. Computer-aided detection of pulmonary nodules based on svm in thoracic ct images. In: *Information and Knowledge Technology (IKT), 2015 7th Conference on*. [S.l.: s.n.], 2015. p. 1–6.
- 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>.
- FAITH, D. P. Phylogenetic pattern and the quantification of organismal biodiversity. *Philosophical Transactions: Biological Sciences*, The Royal Society, v. 345, n. 1311, p. 45–58, 1994. ISSN 09628436. Disponível em: <http://www.jstor.org/stable/56137>.
- 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 International Symposium on Biomedical Imaging: From Nano to Macro, 2011*. [S.l.: s.n.], 2011. p. 169–172. ISSN 1945-7928.
- 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>.
- FIRMINO, M.; ANGELO, G.; MORAIS, H.; DANTAS, M. R.; VALENTIM, R. Computer-aided detection (cade) and diagnosis (cadx) system for lung cancer with likelihood of malignancy. *BioMedical Engineering OnLine*, v. 15, n. 1, p. 1–17, 2016. ISSN 1475-925X. Disponível em: <http://dx.doi.org/10.1186/s12938-015-0120-7>.
- 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>.

GAVRIELIDES, M. A.; KINNARD, L. M.; MYERS, K. J.; PETRICK, N. Noncalcified lung nodules: Volumetric assessment with thoracic ct. *Radiology*, v. 251, n. 1, p. 26–37, 04 2009.

GERLINGER, M.; ROWAN, A. J.; HORSWELL, S.; LARKIN, J.; ENDESFELDER, D.; GRONROOS, E. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, v. 366, n. 10, p. 883–892, 2012. PMID: 22397650. Disponível em: <http://dx.doi.org/10.1056/NEJMoa1113205>.

GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006. ISBN 013168728X.

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](http://dx.doi.org/10.1001/jama.285.7.914).

HADWIGER, H. *Vorlesungen über Inhalt, Oberfläche und Isoperimetrie*. Springer, 1957. (Grundlehren der mathematischen Wissenschaften). Disponível em: <https://books.google.com.br/books?id=-BWoAAAAIAAJ>.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. *Multivariate data analysis*. [S.l.]: Pearson Prentice Hall Upper Saddle River, NJ, 2006. ISBN 0138132631.

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>.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEN, I. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3, n. 6, p. 610–621, Nov 1973. ISSN 0018-9472.

HARDIE, R. C.; ROGERS, S. K.; WILSON, T. A.; 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. Disponível em: <http://dblp.uni-trier.de/db/journals/mia/mia12.html#HardieRWR08>.

HEARD, S. B.; MOOERS, A. O. Phylogenetically patterned speciation rates and extinction risks change the loss of evolutionary history during extinctions. *Proceedings: Biological Sciences*, The Royal Society, v. 267, n. 1443, p. pp. 613–620, 2000. ISSN 09628452. Disponível em: <http://www.jstor.org/stable/2665984>.



HORSTHEMKE, W. H.; RAICU, D. S.; FURST, J. D. Characterizing pulmonary nodule shape using a boundary-region approach. *Medical Imagin: Computer-Aided Diagnosis*, v. 7260, n. 1, p. 72602Y–72602Y–9, 2009. Disponível em: <http://dx.doi.org/10.1117/12.811336>.

INCA. *Estimativas 2014/2015: Incidência de Câncer no Brasil*. 2015. Acesso em: 15 de fevereiro de 2016. Disponível em: <http://www.inca.gov.br/estimativa/2014/>.

IZSAK, J.; PAPP, L. A link between ecological diversity indices and measures of biodiversity. *Ecological Modelling*, v. 130, n. 1â3, p. 151 – 156, 2000. ISSN 0304-3800. Disponível em: <http://www.sciencedirect.com/science/article/pii/S030438000002039>.

IZSAKI, J.; PAPP, L. Application of the quadratic entropy indices for diversity studies of drosophilid assemblages. *Environmental and Ecological Statistics*, Kluwer Academic Publishers, v. 2, n. 3, p. 213–224, 1995. ISSN 1352-8505. Disponível em: <http://dx.doi.org/10.1007/BF00456668>.

JABON, S. A.; RAICU, D. S.; FURST, J. D. Content-based versus semantic-based retrieval: an lfdc case study. *Proc. SPIE*, v. 7263, p. 72631L–72631L–8, 2009. Disponível em: <http://dx.doi.org/10.1117/12.812877>.

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

KEITH, M.; CHIMIMBA, C. T.; REYERS, B.; JAARVELD, A. S. van. Taxonomic and phylogenetic distinctiveness in regional conservation assessments: A case study based on extant south african chiroptera and carnivora. *Animal Conservation*, v. 8, n. 3, p. 279–288, 2005 2005.

KO, J. P.; RUSINEK, H.; JACOBS, E. L.; BABB, J. S.; BETKE, M.; MCGUINNESS, G.; NAIDICH, D. P. Small pulmonary nodules: Volume measurement at chest ct- phantom study. *Radiology*, v. 228, n. 3, p. 864–870, 2003. PMID: 12954901. Disponível em: <http://dx.doi.org/10.1148/radiol.2283020059>.

KREWER, H.; GEIGER, B.; HALL, L.; GOLDFOF, D.; GU, Y.; TOCKMAN, M.; GILLIES, R. Effect of texture features in computer aided diagnosis of pulmonary nodules in low-dose computed tomography. In: *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. [S.l.: s.n.], 2013. p. 3887–3891.

KUHNIGK, J.-M.; DICKEN, V.; BORNEMANN, L.; BAKAI, A.; WORMANN, D.; KRASS, S.; PEITGEN, H. Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic ct scans. *IEEE*

*Transactions on Medical Imaging*, v. 25, n. 4, p. 417–434, April 2006. ISSN 0278-0062.

KUMAR, D.; WONG, A.; CLAUSI, D. A. Lung nodule classification using deep features in ct images. In: IEEE XPLORE. *12th Conference on Computer and Robot Vision (CRV)*. Halifax, NS, Canada: IEEE Xplore, 2015.

KUMAR, S.; RAMESH, J.; VANATHI, P.; GUNAVATHI, K. Robust and automated lung nodule diagnosis from ct images based on fuzzy systems. In: *International Conference on Process Automation, Control and Computing (PACC)*. [S.l.: s.n.], 2011. p. 1–6.

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>.

LEE, S.; KOUZANI, A.; HU, E. Random forest based lung nodule classification aided by clustering. *Computerized Medical Imaging and Graphics*, v. 34, n. 7, p. 535 – 542, 2010. ISSN 0895-6111. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0895611110000418>.

LIU, Y.; YANG, J.; ZHAO, D.; LIU, J. Computer aided detection of lung nodules based on voxel analysis utilizing support vector machines. In: *International Conference on Future BioMedical Information Engineering*. [S.l.: s.n.], 2009. p. 90–93.

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

MECKE, K.; BUCHERT, T.; WAGNER, H. *Robust morphological measures for large-scale structure in the Universe*. Max-Planck-Inst. für Astrophysik, 1993. (Max-Planck-Institut für Astrophysik Garching, München: MPA). Disponível em: <https://books.google.com.br/books?id=jLzgGwAACAAJ>.

MECKE, K.; STOYAN, D. *Morphology of Condensed Matter: Physics and Geometry of Spatially Complex Systems*. Springer, 2002. (Lecture Notes in Physics). ISBN 9783540442035. Disponível em: <https://books.google.com.br/books?id=0BeJY4ZwndsC>.

MICHAL, C.; ELKE, T.; ABHIR, B.; DAVID, P. Integral geometry descriptors for characterizing emphysema and lung fibrosis in hrcr images. *First International Workshop on Pulmonary Image Analysis*, v. 16, p. 155–164, 2008.

- MIYAKE, N.; KIM, H.; ITAI, Y.; TAN, J. K.; ISHIKAWA, S.; KATSURAGAWA, S. Automatic detection of lung nodules in temporal subtraction image by use of shape and density features. In: *Fourth International Conference on Innovative Computing, Information and Control (ICICIC)*. [S.l.: s.n.], 2009. p. 1288–1292.
- MOSMANN, M. P.; BORBA, M. A.; MACEDO, F. P. N. de; LIGUORI, A. d. A. L.; NETO, A. V.; LIMA, K. C. de. Solitary pulmonary nodule and (18)f-fdg pet/ct. part 1: epidemiology, morphological evaluation and cancer probability. *Radiologia Brasileira*, Colégio Brasileiro de Radiologia e Diagnóstico por Imagem, v. 49, n. 1, p. 35–42, Jan-Feb 2016. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4770395/>.
- NASCIMENTO, L. B.; PAIVA, A. C. de; SILVA, A. C. Lung nodules classification in ct images using shannon and simpson diversity indices and svm. In: *Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition*. Berlin, Heidelberg: Springer-Verlag, 2012. (12), p. 454–466. ISBN 978-3-642-31536-7.
- OROZCO, H.; VILLEGAS, O. O. V.; MAYNEZ, L.; SANCHEZ, V.; DOMINGUEZ, H. de J. O. Lung nodule classification in frequency domain using support vector machines. In: *11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*. [S.l.: s.n.], 2012. p. 870–875.
- OROZCO, H. M.; VILLEGAS, O. O. V.; SÁNCHEZ, V. G. C.; DOMÍNGUEZ, H. d. J. O.; ALFARO, M. d. J. N. Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. *BioMedical Engineering OnLine*, BioMed Central, London, v. 14, p. 9, 2015. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4329222/>.
- OROZCO, H. M.; VILLEGAS, O. V.; DOMINGUEZ, H. D. J. O.; SANCHEZ, V. C. Lung nodule classification in ct thorax images using support vector machines. In: *12th Mexican International Conference on Artificial Intelligence (MICAI)*. [S.l.: s.n.], 2013. p. 277–283.
- OST, D.; FEIN, A. M.; FEINSILVER, S. H. The solitary pulmonary nodule. *New England Journal of Medicine*, v. 348, n. 25, p. 2535–2542, 2003. PMID: 12815140. Disponível em: <http://dx.doi.org/10.1056/NEJMcp012290>.
- PARVEEN, S. S.; KAVITHA, C. Article: Classification of lung cancer nodules using svm kernels. *International Journal of Computer Applications*, v. 95, n. 25, p. 25–28, June 2014.
- PATIL, S. S.; GODOY, M. C.; SORENSEN, J. I.; MAROM, E. M. Lung cancer imaging. *Seminars in Diagnostic Pathology*, v. 31, n. 4, p. 293 – 305, 2014. ISSN 0740-2570. Lung Carcinoma: Beyond The who Classification. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0740257014000604>.

- PAVOINE, S.; OLLIER, S.; DUFOUR, A.-B. Is the originality of a species measurable. *Ecology Letters*, Wiley-Blackwell, v. 8, p. 579–586, 2005. Disponível em: [⟨https://hal.archives-ouvertes.fr/hal-00427764⟩](https://hal.archives-ouvertes.fr/hal-00427764).
- PEDRINI, H.; SCHWARTZ, W. R. *Análise de imagens digitais: princípios, algoritmos e aplicações*. [S.l.]: Thomson Learning, 2008. ISBN 8522105952.
- PIANKA, E. R. *Evolutionary Ecology*. 7th ed. ed. Eric R. Pianka, 2011. ISBN 9780321042880. Disponível em: [⟨http://www.amazon.com/Elements-Ecology-EDITION-Robert-Smith/dp/B0047TB21Q⟩](http://www.amazon.com/Elements-Ecology-EDITION-Robert-Smith/dp/B0047TB21Q).
- PIENKOWSKI, M. W.; WATKINSON, A. R.; KERBY, G.; CLARKE, K. R.; WARWICK, R. M. A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology*, Blackwell Science Ltd, v. 35, n. 4, p. 523–531, 1998. ISSN 1365-2664. Disponível em: [⟨http://dx.doi.org/10.1046/j.1365-2664.1998.3540523.x⟩](http://dx.doi.org/10.1046/j.1365-2664.1998.3540523.x).
- POSADAS, P.; ESQUIVEL, D. R. M.; CRISCI, J. V. Using phylogenetic diversity measures to set priorities in conservation: an example from southern south america. *Conservation Biology*, Blackwell Science Inc, v. 15, n. 5, p. 1325–1334, 2001. ISSN 1523-1739. Disponível em: [⟨http://dx.doi.org/10.1111/j.1523-1739.2001.99404.x⟩](http://dx.doi.org/10.1111/j.1523-1739.2001.99404.x).
- RAO, C. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, v. 21, n. 1, p. 24 – 43, 1982. ISSN 0040-5809. Disponível em: [⟨http://www.sciencedirect.com/science/article/pii/0040580982900041⟩](http://www.sciencedirect.com/science/article/pii/0040580982900041).
- REEVES, A.; CHAN, A.; YANKELEVITZ, D.; HENSCHKE, C.; KRESSLER, B.; KOSTIS, W. On measuring the change in size of pulmonary nodules. *IEEE Transactions on Medical Imaging*, v. 25, n. 4, p. 435–450, April 2006. ISSN 0278-0062.
- RIVOLLIER, S.; DEBAYLE, J.; PINOLI, J. C. Shape diagrams for 2d compact sets - part i: analytic convex sets. *Australian Journal of Mathematical Analysis and Applications*, v. 7, n. 2-3, p. 1–27, 2010.
- RIVOLLIER, S.; DEBAYLE, J.; PINOLI, J. C. Shape diagrams for 2d compact sets - part ii: analytic simply connected sets. *Australian Journal of Mathematical Analysis and Applications*, v. 7, n. 2-4, p. 1–21, 2010.
- RODRIGUES, A. S.; GASTON, K. J. Maximising phylogenetic diversity in the selection of networks of conservation areas. *Biological Conservation*, v. 105, n. 1, p. 103 – 111, 2002. ISSN 0006-3207. Disponível em: [⟨http://www.sciencedirect.com/science/article/pii/S0006320701002087⟩](http://www.sciencedirect.com/science/article/pii/S0006320701002087).

- ROTH, K.; BOIKE, J.; VOGEL, H. Quantifying permafrost patterns using minkowski densities. *Permafrost and Periglacial Processes*, John Wiley & Sons, Ltd., v. 16, n. 3, p. 277–290, 2005. ISSN 1099-1530. Disponível em: <http://dx.doi.org/10.1002/ppp.531>.
- 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 Systems with Applications*, v. 42, n. 22, p. 8911 – 8928, 2015. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417415005060>.
- SANTALÓ, L. *Integral Geometry and Geometric Probability*. Beijing World Publishing Corporation (BJWPC), 2004. (Cambridge Mathematical Library). ISBN 9780521523448. Disponível em: <https://books.google.com.br/books?id=xq1iiumncV4C>.
- SARAVANAKUMAR, A.; VAIDEKI, K.; GOVINDARAJAN, K. N.; JAYAKUMAR, S. Establishment of diagnostic reference levels in computed tomography for select procedures in pudhuchery, india. *Journal of Medical Physics / Association of Medical Physicists of India*, Medknow Publications & Media Pvt Ltd, India, v. 39, n. 1, p. 50–55, Jan-Mar 2014. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3931229/>.
- SCHÖLKOPF, B.; SMOLA, A. *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. [S.l.]: MIT Press, 2002.
- SCHUSTER, P. Mathematical modeling of evolution. solved and open problems. *Theory in Biosciences*, v. 130, n. 1, p. 71–89, 2010. ISSN 1611-7530. Disponível em: <http://dx.doi.org/10.1007/s12064-010-0110-z>.
- SCHWEIGER, O.; KLOTZ, S.; DURKA, W.; KÄHN, I. A comparative test of phylogenetic diversity indices. *Oecologia*, Springer-Verlag, v. 157, n. 3, p. 485–495, 2008. ISSN 0029-8549. Disponível em: <http://dx.doi.org/10.1007/s00442-008-1082-2>.
- SCHWEIKERT, G.; ZIEN, A.; ZELLER, G.; BEHR, J.; DIETERICH, C.; ONG, C. S.; PHILIPS, P.; BONA, F. D.; HARTMANN, L.; BOHLEN, A.; KRÜGER, N.; SONNENBURG, S.; RÄTSCH, G. mgene: Accurate svm-based gene finding with an application to nematode genomes. *Genome Research*, Cold Spring Harbor Laboratory Press, v. 19, n. 11, p. 2133–2143, 11 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2775605/>.
- SECHREST, W.; BROOKS, T. M.; FONSECA, G. A. B. d.; KONSTANT, W. R.; MITTERMEIER, R. A.; PURVIS, A.; RYLANDS, A. B.; GITTLEMAN, J. L. Hotspots and the conservation of evolutionary history. *Proceedings of the National Academy of Sciences of the United States of America*, National Academy of

Sciences, v. 99, n. 4, p. pp. 2067–2071, 2002. ISSN 00278424. Disponível em: <http://www.jstor.org/stable/3057919>.

SETIO, A. A. A.; CIOMPI, F.; LITJENS, G.; GERKE, P.; JACOBS, C.; RIEL, S. J. van; WILLE, M. M. W.; NAQIBULLAH, M.; SÁNCHEZ, C. I.; GINNEKEN, B. van. Pulmonary nodule detection in ct images: False positive reduction using multi-view convolutional networks. *IEEE Transactions on Medical Imaging*, v. 35, n. 5, p. 1160–1169, May 2016. ISSN 0278-0062.

SILVA, I. A. da; BATALHA, M. A. Taxonomic distinctness and diversity of a hyperseasonal savanna in central brazil. *Diversity and Distributions*, Blackwell Publishing Ltd, v. 12, n. 6, p. 725–730, 2006. ISSN 1472-4642. Disponível em: <http://dx.doi.org/10.1111/j.1472-4642.2006.00264.x>.

SIVAKUMAR, S.; CHANDRASEKAR, C. Lung nodule detection using fuzzy clustering and support vector machines. *International Journal of Engineering and Technology (IJET)*, v. 5, n. 11, p. 179–185, 2013.

SOLOW, A.; POLASKY, S.; BROADUS, J. On the measurement of biological diversity. *Journal of Environmental Economics and Management*, v. 24, n. 1, p. 60 – 68, 1993. ISSN 0095-0696. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0095069683710041>.

SONE, S.; TAKASHIMA, S.; LI, F.; YANG, Z.; HONDA, T.; MARUYAMA, Y.; HASEGAWA, M.; YAMANDA, T.; KUBO, K.; HANAMURA, K.; ASAKURA, K. Mass screening for lung cancer with mobile spiral computed tomography scanner. *The Lancet*, v. 351, n. 9111, p. 1242 – 1245, 1998. ISSN 0140-6736. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0140673697082299>.

SOUSA, J. R. F. da S.; SILVA, A. C.; PAIVA, A. C. de; NUNES, R. A. Methodology for automatic detection of lung nodules in computerized tomography images. *Computer Methods and Programs in Biomedicine*, v. 98, n. 1, p. 1 – 14, 2010. ISSN 0169-2607. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0169260709002235>.

SPROUFFSKE, K.; ATHENA, A. C.; RADICH, J. P.; CARROLL, M.; NEDELCO, A. M.; MALEY, C. C. An evolutionary explanation for the presence of cancer nonstem cells in neoplasms. *Evolutionary applications*, v. 6, n. 1, p. 92–101, 2013.

STEWART, B. W.; WILD, C. P. *World Cancer Report 2014*. New York: IARC Nonserial Publication, 2014. ISBN 9789283204299.

STRAUSS, S. Y.; WEBB, C. O.; SALAMIN, N. Exotic taxa less related to native species are more invasive. *Proceedings of the National Academy of Science*, v. 103, p. 5841–5845, abr. 2006.

TAN, B. B.; FLAHERTY, K. R.; KAZEROONI, E. A.; IANNETTONI, M. D. The solitary pulmonary nodule\*. *Chest*, v. 123, n. 1, Supplement, p. 89S – 96S, 2003. ISSN 0012-3692. Disponível em: <http://www.sciencedirect.com/science/article/pii/S001236921532986X>.

TARTAR, A.; KILIC, N.; AKAN, A. Classification of pulmonary nodules by using hybrid features. *Computational and Mathematical Methods in Medicine*, Hindawi Publishing Corporation, v. 2013, p. 148363, 2013. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3708407/>.

TERAMOTO, A.; FUJITA, H.; YAMAMURO, O.; TAMAKI, T. Automated detection of pulmonary nodules in pet/ct images: Ensemble false-positive reduction using a convolutional neural network technique. *Medical Physics*, v. 43, n. 6, p. 2821–2827, 2016. Disponível em: <http://scitation.aip.org/content/aapm/journal/medphys/43/6/10.1118/1.4948498>.

UDUPA, J.; HERMAN, G. *3D Imaging in Medicine, Second Edition*. Taylor & Francis, 1999. ISBN 9780849331794. Disponível em: <https://books.google.com.br/books?id=aR6PHYluq4oC>.

VANDAMME, P.; POT, B.; GILLIS, M.; VOS, P. de; KERSTERS, K.; SWINGS, J. Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiological Reviews*, v. 60, n. 2, p. 407–38, 1996. Disponível em: <http://mmbr.asm.org/content/60/2/407.abstract>.

VANE-WRIGHT, R.; HUMPHRIES, C.; WILLIAMS, P. What to protect? - systematics and the agony of choice. *Biological Conservation*, v. 55, n. 3, p. 235 – 254, 1991. ISSN 0006-3207. Disponível em: <http://www.sciencedirect.com/science/article/pii/000632079190030D>.

VAPNIK, V. N.; VAPNIK, V. *Statistical Learning Theory*. [S.l.]: Wiley New York, 1998. ISBN 0471030031.

VARMA, D. R. Managing dicom images: Tips and tricks for the radiologist. *The Indian Journal of Radiology & Imaging*, Medknow Publications & Media Pvt Ltd, India, v. 22, n. 1, p. 4–13, Jan-Mar 2012. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3354356/>.

WAY, T. W.; SAHINER, B.; CHAN, H.-P.; HADJIISKI, L.; CASCADE, P. N.; CHUGHTAI, A.; BOGOT, N.; KAZEROONI, E. Computer-aided diagnosis of pulmonary nodules on ct scans: Improvement of classification performance with nodule surface features. *Medical Physics*, v. 36, n. 7, p. 3086–3098, 2009. Disponível em: <http://scitation.aip.org/content/aapm/journal/medphys/36/7/10.1118/1.3140589>.

- WEBB, C. O.; ACKERLY, D. D.; MCPEEK, M. A.; DONOGHUE, M. J. Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, v. 33, n. 1, p. 475–505, 2002. Disponível em: <http://dx.doi.org/10.1146/annurev.ecolsys.33.010802.150448>).
- WEITZMAN, M. L. On diversity. *The Quarterly Journal of Economics*, JSTOR, v. 107, p. 363–405, 1992. ISSN 00335533.
- YANG, X.; SHEN, X.; LONG, J.; CHEN, H. An improved median-based otsu image thresholding algorithm. *AASRI Procedia*, v. 3, n. 0, p. 468 – 473, 2012. ISSN 2212-6716. Conference on Modelling, Identification and Control. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2212671612002338>).
- YE, X.; LIN, X.; DEHMESHKI, J.; SLABAUGH, G.; BEDDOE, G. Shape-based computer-aided detection of lung nodules in thoracic ct images. *IEEE Transactions on Biomedical Engineering*, v. 56, n. 7, p. 1810–1820, 2009. ISSN 0018-9294.
- YE, X.; LIN, X.; DEHMESHKI, J.; SLABAUGH, G.; BEDDOE, G. Shape-based computer-aided detection of lung nodules in thoracic ct images. *IEEE Transactions on Biomedical Engineering*, v. 56, n. 7, p. 1810–1820, 2009.
- ZHANG, F.; SONG, Y.; CAI, W.; ZHOU, Y.; SHAN, S.; FENG, D. Context curves for classification of lung nodule images. In: *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. [S.l.: s.n.], 2013. p. 1–7.
- 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.