UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

GIOVANNI LUCCA FRANÇA DA SILVA

ANÁLISE DE NÓDULOS PULMONARES USANDO ÍNDICES DE DIVERSIDADES PARA ESTABELECER POSSÍVEIS DIFERENÇAS ENTRE PADRÕES MALIGNOS E BENIGNOS

São Luís
2015

GIOVANNI LUCCA FRANÇA DA SILVA

ANÁLISE DE NÓDULOS PULMONARES USANDO ÍNDICES DE DIVERSIDADES PARA ESTABELECER POSSÍVEIS DIFERENÇAS ENTRE PADRÕES MALIGNOS E BENIGNOS

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

Orientador: Prof. Dr. Aristófanes Corrêa Silva

São Luís
2015

Silva, Giovanni Lucca França da.

Análise de nódulos pulmonares usando índices de diversidades para estabelecer possíveis diferenças entre padrões malignos e benignos/ Giovanni Lucca França da Silva. – São Luís, 2015.

68 f.

Impresso por computador (fotocópia).

Orientador: Aristófanes Corrêa Silva.

Monografia (Graduação) – Universidade Federal do Maranhão, Curso de Ciência da Computação, 2015.

1. Nódulo pulmonar - Diagnóstico. 2. Índice de diversidade taxonômica. 3. Índice de distinção taxonômica. I. Título.

CDU 004.383.5:616.24-006

GIOVANNI LUCCA FRANÇA DA SILVA

ANÁLISE DE NÓDULOS PULMONARES USANDO ÍNDICES DE DIVERSIDADES PARA ESTABELECER POSSÍVEIS DIFERENÇAS ENTRE PADRÕES MALIGNOS E BENIGNOS

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

Aprovada em: 06 / 01 / 2015

BANCA EXAMINADORA

Prof. Dr. Aristófanes Corrêa Silva (Orientador)
Doutor em Informática
Universidade Federal do Maranhão

Prof. Dr. João Dallyson Sousa de Almeida
Doutor em Engenharia de Eletricidade
Universidade Federal do Maranhão

Prof. Ms. Stelmo Magalhães Barros Netto
Mestre em Engenharia de Eletricidade
Universidade Federal do Maranhão

A minha família, em especial aos meus pais e a minha namorada.

AGRADECIMENTOS

A Deus pela saúde, sabedoria e principalmente pelo respirar de cada dia.

A minha família por completa, pais, irmãs, avôs, avós, tios, tias, primos e primas pela educação e carinho ao longo da minha vida.

A minha namorada e futura esposa Yasmim pela compreensão e apoio durante a realização deste trabalho.

A meu orientador Aristófanes, pela paciência, competência e dedicação.

A meus amigos de curso, em especial Caio Eduardo, Caio Belfort, Carlos Augusto, Jefferson, João, Johnatan, Marco e Wendell pelos anos de convivência desde o começo da jornada e aos demais amigos feitos durante o curso.

A meu amigo Oseas, pela ajuda dada ao longo do trabalho, conselhos e sugestões.

A meus amigos do LabPAI, em especial Gilberto, Otílio, Stelmo e Whesley.

A meus amigos de infância, em especial Allan, Isael, Rodrigo, Tassito e Wanderson.

A meus amigos da Igreja Batista Central, em especial ao ministério de louvor e a célula.

A meus amigos do tempo de colégio, em especial Arthur, Caio, Elves, Lucas e Muniz.

À CNPq pelo apoio financeiro.

A todos que de alguma forma contribuíram para minha formação acadêmica.

Fica o meu muito obrigado a todos, que Deus os abençoe!

“Humilhai-vos perante o Senhor, e ele vos exaltará.” Tiago 4:10

RESUMO

O câncer de pulmão é ainda a maior causa de mortalidade por câncer em todo o mundo, com uma das menores taxas de sobrevida a partir do diagnóstico. Por isso, sua detecção precoce é importante para aumentar as chances de cura do paciente, e quanto mais informações o médico dispuser, mais preciso será o diagnóstico. Para auxiliar o especialista na busca e identificação de nódulos e alterações em imagens tomográficas, são desenvolvidos sistemas de detecção assistidos por computador (CAD), que visam automatizar os trabalhos de identificação e classificação dessas estruturas. Diante disso, este trabalho propõe uma metodologia de caracterização de nódulos pulmonares, objetivando tornar-se uma ferramenta computacional utilizada para sugerir sobre a malignidade ou benignidade dos mesmos, atuando como uma segunda opinião junto ao especialista. A metodologia aplicada baseia-se em técnicas de processamento de imagens e reconhecimento de padrões. Os índices de Diversidade Taxonômica (Δ) e de Distinção Taxonômica (Δ^*) foram utilizados como descritores de textura. Os cálculos desses índices são baseados nas árvores filogenéticas, sendo aplicadas neste trabalho na caracterização dos nódulos. Após essa etapa, foi realizada a seleção de características baseada em correlação e a classificação pela Máquina de Vetores de Suporte (MVS), em que foram obtidas taxas de sensibilidade de **86,69%**, especificidade de **90,15%** e acurácia de **89,11%**.

Palavras-chave: Diagnóstico de Nódulo Pulmonar Solitário, Índice de Diversidade Taxonômica, Índice de Distinção Taxonômica e MVS.

ABSTRACT

Lung cancer remains the leading cause of cancer mortality worldwide, with one of the lowest survival rates after diagnosis. Therefore, early detection is important to increase the chances of healing the patient, and the more information the doctor's possession, the more accurate the diagnosis. To assist the user in search and identification of nodules and changes in CT images, detection systems are developed computer aided (CAD), which aim to automate the work of identification and classification of these structures. Thus, this work proposes a methodology for characterizing lung nodules, aiming to become a computational tool used to suggest malignancy or benignity of the same, acting as a second opinion by the expert. The methodology is based on image processing and pattern recognition techniques. Indexes Taxonomic Diversity (Δ) and Taxonomic distinction (Δ *) were used as texture descriptors. The calculations of these indices are based on phylogenetic trees, and are applied in this work on the characterization of nodules. After this stage, the selection of correlation-based features and the classification by Support Vector Machine (SVM) was performed, in which **86.69%** of sensitivity rates were obtained, specificity of **90.15%** and accuracy of **89. 11%.**

Keywords: Diagnosis of Solitary Pulmonary Nodule, Taxonomic Diversity Index, Taxonomic Distinction Index and SVM.

LISTA DE FIGURAS

Figura 1 - Exemplo de Nódulo Pulmonar.....20
Figura 2 - Etapas do Processamento da Imagem Digital.....22
Figura 3 - Exemplo de uma imagem que possui três níveis de cinza (espécies) que é preto, o cinza e o branco. A quantidade de pixels (indivíduos) de preto é 4, de cinza é 2 e de branco é 3. 25
Figura 4 - Representação de uma árvore filogenética para alguns primatas. 26
Figura 5 - Exemplo representando uma imagem em uma árvore taxonômica (à esquerda) e sua matriz de distância (à direita). 27
Figura 6 - Árvore filogenética enraizada na forma de cladograma inclinado. 27
Figura 7 - Separação entre duas classes através de hiperplanos.....31
Figura 8 - Vetores de Suporte (destacado por círculos). 33
Figura 9 - Etapas da Metodologia.....36
Figura 10 - Ilustração do resumo das marcações dos nódulos**.** 37
Figura 11 - Ilustração do resumo do diagnóstico dos nódulos. 38
Figura 12 - Procedimento da criação das cinco máscaras internas.....40
Figura 13 - Procedimento da criação das quatro máscaras externas. 40
Figura 14 - Árvore 1: Árvore enraizada na forma de cladograma inclinado.....41
Figura 15 - Descrição da quantidade de arestas das espécies 0 com 1 (a), 0 com 2 (b) e 1 com 3 (c).....42
Figura 16 - Árvore 2: Modelo criado a partir da Árvore 1, com eliminação das espécies sem indivíduos. 42
Figura 17 - Descrição da quantidade de arestas das espécies 0 com 2 (a), 0 com 3 (b) e 1 com 3 (c).....43

LISTA DE TABELAS

Tabela 1 - Resultados obtidos pela Árvore 1 para as 3 Classes. 46
Tabela 2 - Resultados obtidos pela Árvore 1 para as 3 Classes com Seleção de Características. 47
Tabela 3 - Resultados obtidos pela Árvore 2 para as 3 Classes. 48
Tabela 4 - Resultados obtidos pela Árvore 2 para as 3 Classes com Seleção de Características. 48
Tabela 5 - Resultados obtidos pela Árvore 3 para as 3 Classes. 49
Tabela 6 - Resultados obtidos pela Árvore 3 para as 3 Classes com Seleção de Características. 50
Tabela 7 - Resultados obtidos pela Junção de todas as Árvores para as 3 Classes. 50
Tabela 8 - Resultados obtidos pela Junção de todas as Árvores para as 3 Classes com Seleção de Características. 51
Tabela 9 - Resultados obtidos pela Árvore 1 para as 2 Classes. 53
Tabela 10 - Resultados das médias das proporções obtidas pela Árvore 1. 53
Tabela 11 - Resultados obtidos pela Árvore 1 para as 2 Classes com Seleção de Características. 54
Tabela 12 - Resultados das médias das proporções obtidas pela Árvore 1 com Seleção de Características. 54
Tabela 13 - Resultados obtidos pela Árvore 2 para as 2 Classes. 55
Tabela 14 - Resultados das médias das proporções obtidas pela Árvore 2. 55
Tabela 15 - Resultados obtidos pela Árvore 2 para as 2 Classes com Seleção de Características. 56
Tabela 16 - Resultados das médias das proporções obtidas pela Árvore 2 com Seleção de Características. 56
Tabela 17 - Resultados obtidos pela Árvore 3 para as 2 Classes. 57
Tabela 18 - Resultados das médias das proporções obtidas pela Árvore 3. 57
Tabela 19 - Resultados obtidos pela Árvore 3 para as 2 Classes com Seleção de Características. 58
Tabela 20 - Resultados das médias das proporções obtidas pela Árvore 3 com Seleção de Características. 58
Tabela 21 - Resultados obtidos pela Junção de todas as Árvores para as 2 Classes. 59
Tabela 22 - Resultados das médias das proporções obtidas pela Junção de todas as Árvores. 60
Tabela 23 - Resultados obtidos pela Junção de todas as Árvores para as 2 Classes com Seleção de Características. 60
Tabela 24 - Resultados das médias das proporções obtidas pela Junção de todas as Árvores com Seleção de Características 61
Tabela 25 - Melhores e piores resultados obtidos com a metodologia proposta. 62
Tabela 26 - Comparação dos resultados entre trabalhos relacionados. 62

LISTA DE ABREVIATURAS

AC	Acurácia
CAD	*Computer-Aided Detection*
CADx	*Computer-Aided Detection and Diagnosis*
CFS	*Correlation-based Feature Selection*
DICOM	*Digital Imaging and Communications in Medicine*
ES	Especificidade
EUA	Estados Unidos da América
FN	Falso Negativo
FP	Falso Positivo
INCA	Instituto Nacional do Câncer José Alencar Gomes da Silva
LIDC	*Lung Image Database Consortium*
LIDC-IDRI	*Lung Image Database Consortium - Image Database Resource Initiative*
ME1	Máscara Externa 1
ME2	Máscara Externa 2
ME3	Máscara Externa 3
ME4	Máscara Externa 4
MI1	Máscara Interna 1
MI2	Máscara Interna 2
MI3	Máscara Interna 3
MI4	Máscara Interna 4
MI5	Máscara Interna 5
MVS	Máquina de Vetores de Suporte
NCI	*National Cancer Institute*
Q8	Quantizada em 8 bits
Q12	Quantizada em 12 bits
Q16	Quantizada em 16 bits
ROI	*Region of Interest*
SE	Sensibilidade
TC	Tomografia Computadorizada
UH	Unidade de Hounsfield
VN	Verdadeiro Negativo
Voxel	*Volume Element*
VP	Verdadeiro Positivo
XML	*eXtensible Markup Language*
3D	Tri-dimensional
Δ	Índice de Diversidade Taxonômica
Δ^*	Índice de Distinção Taxonômica

SUMÁRIO

1. INTRODUÇÃO 15
1.1 Motivação 16
1.2 Objetivos 16
1.3 Trabalhos Relacionados 16
1.4 Organização do Trabalho 19
2. FUNDAMENTAÇÃO TEÓRICA 20
2.1 Nódulo Pulmonar Solitário 20
2.2 Tomografia Computadorizada 21
2.3 Técnicas de Processamento de Imagens 22
2.3.1 Quantização Uniforme 23
2.4 Análise de Textura 24
2.5 Índice de Diversidade 25
2.5.1 Diversidade Filogenética 26
2.5.2 Índices Taxonômicos 28
2.6 Seleção de Características 29
2.6.1 Seleção de Características baseada em Correlação 29
2.7 Reconhecimento de Padrões 30
2.7.1 Máquina de Vetores de Suporte 30
2.8 Validação dos Resultados 34
3. METODOLOGIA 36
3.1 Aquisição das Imagens 36
3.2 Segmentação 3D dos Nódulos 39
3.3 Extração de Características 39
3.3.1 Abordagem com Máscara Interna 39
3.3.2 Abordagem com Máscara Externa 40
3.3.3 Árvore 1 – Árvore Enraizada na Forma de Cladograma Inclinado 41
3.3.4 Árvore 2 – Árvore Enraizada na Forma de Cladograma Inclinado Excluindo as Espécies sem Indivíduos 42
3.3.5 Árvore 3 – Árvore Enraizada na Forma de Cladograma Inclinado Modificando as Arestas 43
3.4 Seleção de Características 43
3.5 Reconhecimento de Padrões 43

3.6 Validação dos Resultados 44
4. RESULTADOS E DISCUSSÃO 45
4.1 Testes com as Classes Benigno, Maligno e Indeterminado 46
4.1.1 Árvore 1 46
4.1.2 Árvore 2 47
4.1.3 Árvore 3 49
4.1.4 Todas as Árvores Juntas 50
4.2 Testes com as Classes Benigno e Maligno 52
4.2.1 Árvore 1 52
4.2.2 Árvore 2 54
4.2.3 Árvore 3 57
4.2.4 Todas as Árvores Juntas 59
5. CONCLUSÃO 63
REFERÊNCIAS 65

1. INTRODUÇÃO

O câncer é o nome dado a um conjunto de mais de 100 doenças que têm em comum o crescimento desordenado (maligno) de células que invadem os tecidos e órgãos, podendo espalhar-se (metástase) para outras regiões do corpo. Dividindo-se rapidamente, estas células tendem a ser muito agressivas e incontroláveis, determinando a formação de tumores (acúmulo de células cancerosas) ou neoplasias malignas. Por outro lado, um tumor benigno significa simplesmente uma massa localizada de células que se multiplicam vagarosamente e se assemelham ao seu tecido original, raramente constituindo um risco de vida. As causas de 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. Esses fatores causais podem interagir de várias formas, aumentando a probabilidade de transformações malignas nas células normais (INCA, 2014).

O câncer de pulmão é o mais frequente de todos os tumores malignos, apresentando aumento de 2% por ano na sua incidência mundial. A última estimativa mundial apontou incidência de 1,82 milhões de casos novos de câncer de pulmão para o ano de 2012, sendo 1,24 milhões em homens e 583 mil em mulheres. Em 90% dos casos diagnosticados, o câncer de pulmão está associado ao consumo de derivados de tabaco. No Brasil, foi responsável por 22.424 mortes em 2011. Altamente letal, a sobrevida média cumulativa total em cinco anos varia entre 13 e 21% em países desenvolvidos e entre 7 e 10% nos países em desenvolvimento. No fim do século XX, o câncer de pulmão se tornou uma das principais causas de morte evitáveis. No Brasil, as estimativas de casos de câncer de pulmão do ano de 2014 foram de 27.330, sendo 16.400 homens e 10.930, mulheres (INCA, 2014).

A maneira mais fácil de diagnosticar o câncer de pulmão é por meio de raios-X do tórax, complementado por uma tomografia computadorizada (TC). Essas tecnologias de aquisição de imagens auxiliam os especialistas. Por meio das imagens obtidas, é possível desenvolver sistemas assistidos por computador que auxiliam na detecção do nódulo, e que podem até fornecer uma possível indicação de diagnóstico, funcionando assim com uma segunda opinião para análise dos exames. Esses sistemas podem ser do tipo CAD (*Computer-*

Aided Detection – Detecção assistida por computador) ou CADx (*Computer-Aided Detection and Diagnosis* – Detecção e Diagnóstico assistida por computador).

Os sistemas CAD auxiliam os especialistas na detecção de regiões de interesse no exame, mas não realizam o diagnóstico. Os sistemas CADx sugerem um diagnóstico (maligno ou benigno, por exemplo). Sistemas como CADx utilizam técnicas de processamento de imagens para auxiliar na realização do diagnóstico.

1.1 Motivação

O câncer de pulmão é ainda a maior causa de mortalidade por câncer em todo o mundo, com uma das menores taxas de sobrevida a partir do diagnóstico. Por essas razões, nas últimas décadas tem surgido um grande interesse no desenvolvimento e uso de técnicas de processamento de imagens digitais de tomografia computadorizada com o objetivo de auxiliar o diagnóstico dos pacientes. Assim, quanto mais precoce o diagnóstico, mais chances de cura terão os pacientes e o quão maior o número de informações à disposição do especialista, mais preciso será o diagnóstico.

1.2 Objetivos

Este trabalho tem como principal objetivo propor uma metodologia para diagnóstico de nódulos pulmonares, por meio da análise de imagens de tomografia computadorizada, de forma a determinar padrões de comportamento de nódulos benignos e malignos, utilizando medidas de textura obtidas com o cálculo de índices de diversidade: Índice de Diversidade Taxonômica e de Distinção Taxonômica. A metodologia pode ainda ser incorporada a um sistema do tipo CADx, não substituindo a função do especialista, e sim oferecendo a ele uma segunda opinião, corroborando para o aumento da produtividade e melhoria nas taxas de diagnósticos corretos.

1.3 Trabalhos Relacionados

Muitas metodologias computacionais têm sido desenvolvidas para a tarefa de detecção e diagnóstico a partir de imagens de tomografia computadorizada. Em Silva (2009) apresenta

uma metodologia para diagnóstico de nódulos pulmonares solitários em maligno e benigno utilizando a base de dados *Lung Image Database Consortium* (LIDC). A representação dos nódulos foi feita com a extração de medidas de geometria e de textura sendo esta última através do Índice de Diversidade de Shannon. As medidas foram submetidas para classificação na Máquina de Vetores de Suporte (MVS) obtendo taxas de sensibilidade de 90%, especificidade de 96,67 e acurácia de 95%.

Em Kumar *et al*. (2011) apresenta um sistema que pode detectar e diagnosticar nódulos pulmonares em malignos ou benigno em tomografia computadorizada de pulmão. Na primeira fase do sistema, a imagem de entrada é pré-processada e a região do nódulo é segmentada. A segunda fase inclui o diagnóstico do nódulo baseado em sistema *fuzzy*, que por sua vez é baseado na área e no nível de cinza da região do nódulo. Utilizando uma base proprietária com 40 casos, o método alcança bons resultados com acurácia de 90%.

Em Nascimento (2012) apresenta um método para diagnóstico de nódulos pulmonares em maligno e benigno, através de tomografia computadorizada utilizando medidas de textura obtidas com os cálculos dos índices de diversidade de Shannon e de Simpson e a Máquina de Vetores de Suporte para a classificação, obtendo taxas de sensibilidade de 85,64%, especificidade de 97,89% e acurácia de 92,78% para a base de dados *Lung Image Database Consortium* (LIDC) e taxas de sensibilidade de 82,95%, especificidade de 84,58% e acurácia de 83,75% para a base de dados *Lung Image Database Consortium - Image Database Resource Initiative* (LIDC-IDRI).

Em Oliveira (2013) apresenta uma metodologia para avaliar o desempenho dos índices taxonômicos, índice de diversidade taxonômica e índice de distinção taxonômica, a partir da geração de modelos de árvores filogenéticas, como método de extração de características de textura de regiões de interesse em imagens mamográficas e depois discriminar as regiões em massa e não massa, alcançando acurácia média de 99,67%, especificidade média de 99,25% e sensibilidade média de 100%.

Em Carvalho (2014) apresenta uma metodologia baseada em técnicas de processamento de imagens e reconhecimento de padrões, utilizando os índices de diversidade taxonômica e distinção taxonômica para descrição da textura dos candidatos a nódulos e não-

nódulos, juntamente com arquiteturas de árvores filogenéticas obtendo acurácia média de 99,22%, sensibilidade média de 98% e especificidade média de 97%.

Em Parveen & Kavitha (2014) apresenta uma metodologia para classificação de nódulos pulmonares usando a Máquina de Vetores de Suporte com diferentes funções de núcleo. Para a extração de características, utilizou-se a matriz de co-ocorrência de níveis de cinza (GLCM) como descritor da textura. A metodologia alcançou taxas de sensibilidade de 83,45% e de especificidade de 82,23% usando o núcleo linear, taxas de sensibilidade de 85,79% e de especificidade de 84,91% usando o núcleo polinomial e taxas de sensibilidade de 91,38% e de especificidade de 89,56% usando o núcleo função de base radial.

Em Freire (2014) apresenta uma metodologia para a análise do uso de medidas de diversidade como descritores de textura de estruturas internas ao parênquima pulmonar, com a finalidade de classificá-las em nódulos e não-nódulos. Uma comparação com medidas geométricas foi feita a fim de validar a eficiência das medidas propostas. A metodologia alcançou sensibilidade de 90,45%, com 92,51% de acurácia e valor de área da curva ROC de 0,897 utilizando somente as medidas de diversidade, e sensibilidade de 92,75%, com 93,21% de acurácia e valor de área da curva ROC de 0,904 utilizando as medidas de diversidade e geometria.

Em Tartar, Akan & Kilic (2014) apresenta um novo sistema de detecção e diagnóstico assistida por computador (CADx) para a classificação de nódulos pulmonares em maligno e benigno. O sistema CADx proposto utiliza classificadores de aprendizagem *ensemble*. A metodologia utilizando classificadores *bagging* obteve taxas de sensibilidade de 94,7%, 90% e 77,8% para as classes benigno, maligno e indeterminado, respectivamente.

Vários trabalhos relacionados apresentam boas taxas de acurácia para o diagnóstico de nódulos pulmonares em imagens de tomografia computadorizada. No entanto, ainda é necessário identificar técnicas que permitam melhorar e consolidar estes resultados. Verifica-se que a classificação de nódulos pulmonares quanto a sua malignidade e benignidade é ainda um problema em aberto, e que medidas de textura se mostram muito promissoras para essa discriminação.

1.4 Organização do Trabalho

O restante deste trabalho está constituído de mais quatro capítulos, descritos resumidamente a seguir.

No Capítulo 2 é exposta a fundamentação teórica necessária para a compreensão da metodologia proposta. Neste capítulo são descritos conceitos básicos sobre o nódulo pulmonar solitário, a tomografia computadorizada, a técnica de processamento de imagens digitais quantização uniforme, a extração de textura através dos Índices Taxonômicos (Diversidade e Distinção), a técnica de seleção de características baseada em correlação, a técnica de reconhecimento de padrões denominada Máquina de Vetores de Suporte e as métricas de validação dos resultados.

No Capítulo 3 é descrita a metodologia para realizar a classificação dos nódulos pulmonares em maligno e benigno, extraídos dos exames de tomografias computadorizadas, a aquisição das imagens, a segmentação dos nódulos, a extração das características de textura baseada nos índices taxonômicos, a seleção de características, a classificação através da máquina de vetores de suporte e a validação dos resultados.

No Capítulo 4 são expostos e discutidos os resultados alcançados por meio da metodologia. Por fim, o Capítulo 5 apresenta a conclusão deste trabalho, mostrando a eficiência dos métodos utilizados e oferecendo sugestões para trabalhos futuros.

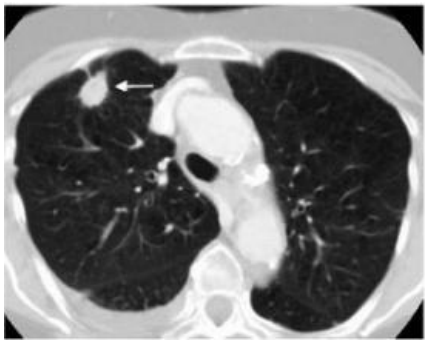
2. FUNDAMENTAÇÃO TEÓRICA

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

2.1 Nódulo Pulmonar Solitário

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

Figura 1 - Exemplo de Nódulo Pulmonar.



Fonte: (ARMATOIII *et al*.,2011).

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

a) Tamanho: Os dados da literatura demonstram claramente que a probabilidade de malignidade aumenta com o aumento do tamanho do nódulo pulmonar.

b) Localização: Os cânceres de pulmão são mais frequentes nos lobos superiores.

c) Existência ou Não de Calcificação e Gordura: A existência de calcificação constitui evidência quase certa de benignidade com raríssimas exceções. A identificação de gordura no interior de um nódulo pulmonar também é uma característica fortemente sugestiva de benignidade quase sem exceções.

d) Tempo de Duplicação: Um nódulo com tempo de duplicação muito curto (por exemplo, inferior a um mês) ou muito longo (classicamente, superior a 450 dias) tem maior

probabilidade de ser benigno. Ao contrário, um nódulo pulmonar cujo tempo de duplicação estiver entre esses limites tem maior chance de revelar-se maligno.

Todavia, o diagnóstico definitivo de malignidade é dado somente pelo exame citopatológico[1] do material obtido por procedimentos que estão se tornando de menor morbidade, como a biopsia transbrônquica e transtorácica.

2.2 Tomografia Computadorizada

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

A tomografia computadorizada baseia-se nos mesmos princípios técnicos que a radiografia tradicional, e na verdade é uma evolução técnica dela, que usa uma radiação maior e toma imagens fatiadas dos segmentos que examina, as quais o médico superpõe imaginativamente para obter uma visão tridimensional. Em alguns casos há necessidade de se utilizar um contraste injetável, a fim de aumentar a capacidade diagnóstica. As imagens da tomografia podem ser tomadas em dois planos básicos: o axial (perpendicular ao maior eixo do corpo) e o coronal (paralelo à sutura coronal do crânio) e permitem reconstruções no plano sagital (paralelo à sutura sagital do crânio) e tridimensionais (ABC MED, 2014).

A tomografia computadorizada é usada para detectar tumores, fraturas, obstruções circulatórias, alterações nas estruturas orgânicas e outras anomalias teciduais, sendo mais precisa para tecidos moles que as simples radiografias. Hoje em dia a tomografia computadorizada vem apresentando menor volume de exames em comparado a ressonância magnética em virtude de duas grandes vantagens dessa última: imagens com maior definição e o fato de não usar energia radioativa (ABC MED, 2014).

[1] Exame feito para estudar as alterações celulares, principalmente as do núcleo das células.

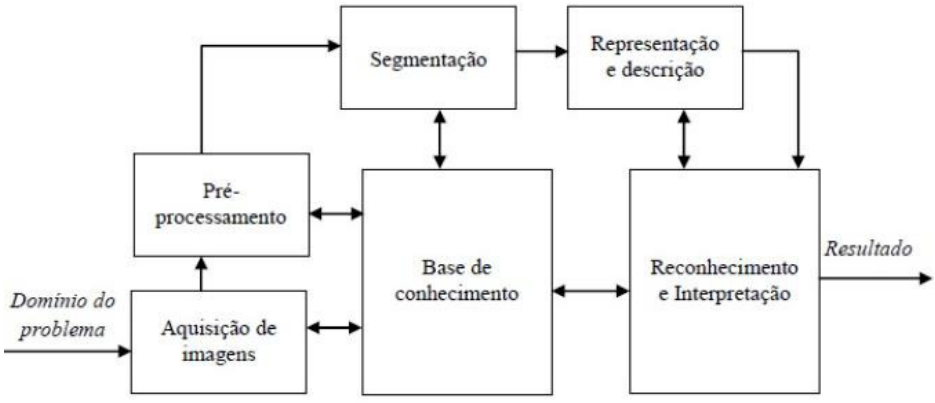
2.3 Técnicas de Processamento de Imagens

A formação de uma imagem digital é definida como uma função bidimensional, em que x e y são coordenadas espaciais e a amplitude de *f* em um par de coordenadas (x,y) é denominada intensidade ou nível de cinza da imagem naquele ponto (GONZALEZ & WOODS, 2002). Em imagens tridimensionais (3D), como as imagens de tomografia computadorizadas utilizadas neste trabalho, essa representação elementar do ponto, agora com coordenadas espaciais x, y e z, é chamada de *voxel* (*volume element*) (FACON, 2002).

O processamento de imagens digitais compreende processos cujas entradas e saídas são imagens e, além disso, engloba os processos de extração de características a partir de imagens, incluindo o reconhecimento de objetos individuais (GONZALEZ & WOODS, 2002). Um dos objetivos principais desse processamento é melhorar a informação visual para interpretação humana e os dados para percepção automática através de máquinas (GONZALEZ & WOODS, 2002).

A Figura 2 apresenta um esquema utilizado para demonstrar as diversas fases do processamento da imagem. Após o total domínio do problema. Seguem-se as etapas: aquisição das imagens digitais, pré-processamento, segmentação, representação e descrição, reconhecimento e interpretação. O conjunto de resultados gerados por uma etapa é utilizado pela próxima. Nem sempre esse conjunto gerado é uma imagem. Sendo que ao fim de todas as etapas, o resultado pode ser ou não representado por uma imagem digital.

Figura 2 - Etapas do Processamento da Imagem Digital.



Fonte: (GONZALES & WOODS, 2002).

O primeiro passo então é a aquisição da imagem, isto é, procedimento em que um digitalizador converte a imagem analógica para digital. Neste trabalho é utilizada a base de dados pública de exames de tomografia computadorizada *Lung Image Database Consortium - Image Database Resource Initiative* (LIDC-IDRI) obtidas por meio do tomógrafo.

A segunda etapa é o pré-processamento das imagens adquiridas. Esta etapa tem como finalidade melhorar certas partes da imagem para aumentar as chances de sucesso dos processos seguintes. Este trabalho não fez uso de nenhuma técnica de pré-processamento para a segmentação dos nódulos das imagens de tomografia computadorizada.

A terceira etapa é a segmentação que tem como finalidade extrair das imagens apenas as partes que realmente interessam para o processamento. Neste trabalho a segmentação é definida por especialistas que delimitam manualmente os nódulos pulmonares.

A quarta etapa é a representação e descrição, também conhecida como extração de características. Essa etapa tem como finalidade determinar as características que resultam em informação quantitativa de interesse ou que sejam básicas para discriminação entre classes distintas. Na extração de características fez-se uso dos índices de Diversidade Taxonômica (Δ) e de Distinção Taxonômica (Δ^*) para construir o vetor de características que representa o nódulo pulmonar.

A quinta e última etapa é o reconhecimento e interpretação das imagens. O reconhecimento é responsável por atribuir um rótulo a um objeto, baseado em suas características, enquanto a interpretação atribui um significado a um conjunto de objetos reconhecidos. Para este trabalho um rótulo previamente determinado indica no resultado do processo, qual a natureza dos nódulos submetidos à classificação.

2.3.1 Quantização Uniforme

Uma imagem digital é discretizada espacialmente em x e y, e também em amplitude (intensidade luminosa). A discretização em amplitude é conhecida como quantização, e a outra, denominada amostragem (GONZALEZ & WOODS, 2002).

A quantização uniforme consiste em dividir a escala de cinza da imagem em intervalos iguais, em que cada intervalo é mapeado para um valor de cinza na imagem quantizada, de modo que a escala de cinza da imagem quantizada é dada por [0, L'- 1], sendo o nível de cinza da imagem quantizada (L') menor do que da imagem original (L), ou seja, L' < L (PEDRINI & SCHWARTZ, 2008).

A expressão para calcular esse mapeamento é:

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

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

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

2.4 Análise de Textura

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

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

Neste trabalho foi proposta a descrição da textura dos tecidos de regiões (Seção 3.3.1 e Seção 3.3.2) dos nódulos pulmonares solitários através do Índice de Diversidade Taxonômica (Δ) e do Índice de Distinção Taxonômica (Δ*), que são medidas estatísticas.

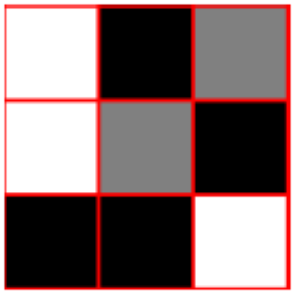
2.5 Índice de Diversidade

A diversidade é um termo muito utilizado na área da ecologia. O seu objetivo é informar a variedade de espécies presentes em uma comunidade ou área. O conceito de comunidade é descrito como um conjunto de espécies que ocorrem em um determinado lugar e tempo (MAGURRAN, 2004). 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 (GIBBONS, KOTZ & JOHNSON, 1988).

Para medir a diversidade temos duas componentes: a riqueza de espécies, que consiste no número de espécies encontradas em determinada região, e a abundância relativa, que é o número de indivíduos de uma determinada espécie existentes numa dada área (PIANKA, 1994). O resultado do cálculo, para qualquer índice de diversidade, é representado por um único valor (SANTOS, 2009). As medidas de diversidade de espécies são geralmente úteis para comparar padrões em diferentes áreas.

A forma mais simples da aplicação do índice de diversidade em imagens é quando a comunidade representa uma imagem ou região da mesma, as espécies sendo os níveis de cinza e os indivíduos sendo os pixels (SOUSA, 2011). A Figura 3 mostra um exemplo demonstrativo.

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

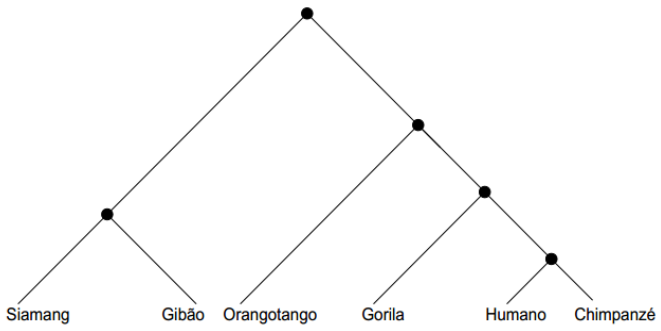


Fonte: (OLIVEIRA, 2013).

2.5.1 Diversidade Filogenética

A filogenia é um ramo da biologia responsável pelo estudo das relações evolutivas entre as espécies, pela verificação dos relacionamentos entre elas, a fim de determinar possíveis ancestrais comuns. Uma árvore filogenética, ou simplesmente filogenia, é uma árvore onde as folhas representam os organismos e os nós internos representam supostos ancestrais. As arestas da árvore denotam as relações evolutivas (ARAÚJO, 2003). Na Figura 4 temos um exemplo de árvore filogenética, em que se verifica o relacionamento entre espécies de macacos e a espécie humana, onde podemos ver que o homem e o chimpanzé são geneticamente mais próximos que os outros pares presentes na árvore (ARAÚJO, 2003).

Figura 4 - Representação de uma árvore filogenética para alguns primatas.



Fonte: (ARAÚJO, 2013).

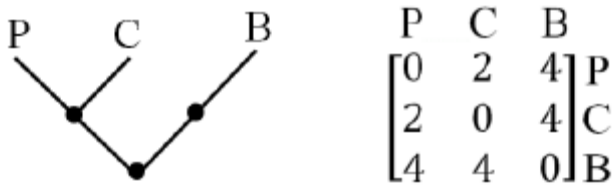
De maneira geral, a diversidade não pode ser medida apenas com a utilização de dados como a abundância e a riqueza de espécies, cada vez mais o parâmetro filogenético vem sendo inserido neste cálculo (CLARKE & WARWICK, 1998). A diversidade filogenética é uma medida da diversidade de uma comunidade que incorpora as relações filogenéticas das espécies (MAGURRAN, 2004). A combinação da abundância das espécies com a proximidade filogenética para gerar um índice de diversidade é denotada diversidade taxonômica (SILVA & BATALHA, 2006). A taxonomia é a ciência que lida com a classificação (criação de novas taxas), identificação (alocação de linhagens dentro de espécies conhecidas) e nomenclatura (VANDAMME *et al.*, 1996).

Clarke e Warwick (1998) desenvolveram um método para mensurar a diversidade taxonômica muito sensível a perturbações ambientais e apropriado para avaliar as diferenças entre comunidades. Uma comunidade em que as espécies estão distribuídas em muitos

gêneros deve apresentar uma diversidade maior que uma comunidade em que a maioria das espécies pertence a um mesmo gênero (MAGURRAN, 2004).

A diversidade taxonômica é baseada no conjunto das distâncias entre pares de espécies acumuladas a partir das árvores taxonômicas (RICOTTA, 2004). A Figura 5 apresenta uma ilustração de uma árvore taxonômica em que as folhas são espécies e a soma da quantidade de arestas que ligam determinado par de espécies é informada pela matriz.

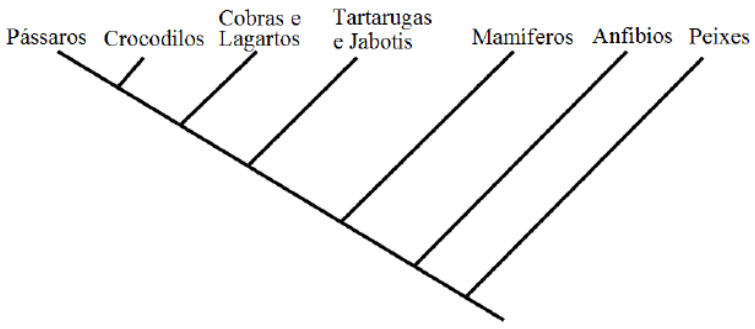
Figura 5 - Exemplo representando uma imagem em uma árvore taxonômica (à esquerda) e sua matriz de distância (à direita).



Fonte: (OLIVEIRA, 2013).

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

Figura 6 - Árvore filogenética enraizada na forma de cladograma inclinado.



Fonte: (VIANA, 2007).

2.5.2 Índices Taxonômicos

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

O Índice de Diversidade Taxonômica (Δ) considera a abundância das espécies e a relação taxonômica entre elas, assim, o seu valor expressa a distância taxonômica média entre quaisquer dois indivíduos, escolhidos na amostra ao acaso (GORENSTEIN, 2009).

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

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

Já o Índice de Distinção Taxonômica (Δ*) representa a distância taxonômica média entre dois indivíduos, com a restrição de que sejam de espécies diferentes (GORENSTEIN, 2009).

$$\Delta^* = \frac{\sum\sum_{i<j} \omega_{ij} x_i x_j}{\sum\sum_{i<j} x_i x_j} \quad (3)$$

Baseado na Figura 3 e Figura 5, a aplicação dos índices Δ e Δ* é demonstrado a seguir:

$$\Delta = \frac{\omega_{pc} x_p x_c + \omega_{pb} x_p x_b + \omega_{cb} x_c x_b}{\frac{n(n-1)}{2}} = \frac{2.4.2 + 4.4.3 + 4.2.3}{\frac{3(3-1)}{2}} = 29{,}3$$

$$\Delta^* = \frac{\omega_{pc} x_p x_c + \omega_{pb} x_p x_b + \omega_{cb} x_c x_b}{x_p x_c + x_p x_b + x_c x_b} = \frac{2.4.2 + 4.4.3 + 4.2.3}{4.2 + 4.3 + 2.3} = 3{,}38$$

2.6 Seleção de Características

Comumente, um problema em aplicações de visão computacional é a utilização de grande número de características. Embora, intuitivamente, quanto maior for esse número maior o poder discriminatório do classificador, de maneira geral, nem todas as características são necessárias para discriminar as classes de maneira precisa e incluí-las no modelo de classificação pode até mesmo gerar resultados inferiores do que seriam obtidos se elas fossem removidas (PAPPA, 2002). Características irrelevantes ou redundantes podem confundir o algoritmo de aprendizagem, ajudando a esconder as distribuições de pequenos conjuntos de características realmente relevantes (KOLLER & SAHAMI, 1996).

2.6.1 Seleção de Características baseada em Correlação

A técnica de seleção de características *Correlation-based Feature Selection* (CFS) parte da hipótese de que um bom subconjunto de características é aquele que contêm características com alta correlação com a classe, mas com uma baixa correlação entre si (HALL & SMITH, 1998).

Dessa forma, o CFS avalia o subconjunto de características considerando a capacidade preditiva individual de cada um em conjunto com o grau de redundância entre eles. Subconjuntos de características que são muito correlacionados com a classe ao mesmo tempo em que possuem uma baixa correlação entre si são preferidos para a seleção.

Primeiramente o CFS calcula uma matriz de correlação de característica-classe e característica-característica. E após isso, a relevância de um subconjunto de características pode ser definida como:

$$CFS(X_k) = \frac{k * \overline{r_{kc}}}{\sqrt{k + k(k-1) * \overline{r_{kk}}}} \quad (4)$$

onde $CFS(X_k)$ é a heurística "mérito" de um subconjunto de características X contendo k características, $\overline{r_{kc}}$ é a média da correlação entre característica-classe e $\overline{r_{kk}}$ é a média da correlação entre característica-característica.

O numerador da Equação 4 pode ser visto como um indicador do poder preditivo do conjunto de características e o denominador da Equação 4 indica o grau de redundância que existe entre as características. O CFS começa com o conjunto vazio de características e usa a heurística *best-first-search* com um critério de parada de 5 consecutivos subconjuntos que não melhoram o "mérito". O subconjunto com o maior "mérito" encontrado pela heurística será o subconjunto selecionado.

A técnica de seleção de características baseada em correlação é aplicada nesse trabalho, após a extração de características, para determinar as características que melhor discriminam os nódulos quanto à classe (maligno e benigno).

2.7 Reconhecimento de Padrões

O reconhecimento de padrões por computador é uma das mais importantes ferramentas usadas no campo da inteligência de máquina. Atualmente está presente em inúmeras áreas do conhecimento, encontrando aplicações diretas e visão computacional, na análise sísmica, no reconhecimento de voz, no reconhecimento de faces, na identificação de íris, na identificação de digitais, no reconhecimento de caracteres impressos e manuscritos, além de outras (NOGUEIRA, 2007).

No presente trabalho foi utilizada uma técnica de aprendizado de máquina para o reconhecimento de padrões, denominada Máquina de Vetores de Suporte (MVS).

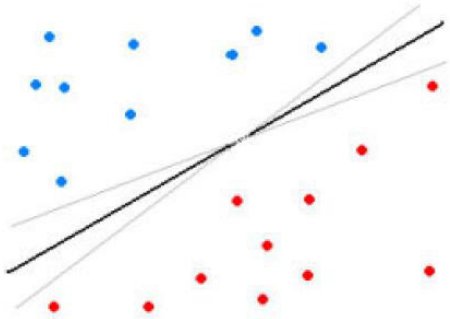
2.7.1 Máquina de Vetores de Suporte

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

A Figura 7 mostra em duas dimensões, para melhor visualização, hiperplanos de separação entre duas classes linearmente separáveis. O hiperplano ótimo (linha mais escura), não somente separa as duas classes, mas mantém a maior distância possível com relação aos pontos da amostra.

Há casos em que podem existir vários possíveis hiperplanos de separação, mas MVS busca apenas encontrar o que maximize a margem entre os exemplos de treinamento (BRAGA, 2005).

Figura 7 - Separação entre duas classes através de hiperplanos.



Fonte: (NASCIMENTO, 2012).

Seja o conjunto de amostras de treinamento (x_i, y_i) sendo, $x_i \in \mathbb{R}^n$ o vetor de entrada, y_i a classificação correta das amostras e $i = 1, 2,..., n$ o índice de cada ponto amostral. O objetivo da classificação é estimar a função $f(x)$: $\mathbb{R}^n \rightarrow \{\pm 1\}$ que separe corretamente os exemplos de teste em classes distintas.

A etapa de treinamento estima a função $f(x) = (w.x) + b$, procurando valores tais que a seguinte relação seja satisfeita:

$$y_i\big((w.x_i) + b\big) \geq 1 \qquad (5)$$

Sendo w o vetor normal ao hiperplano de decisão e b o corte ou distância da função f em relação à origem, os valores ótimos de w e b serão encontrados de acordo com a restrição dada pela Equação 5 ao minimizar a seguinte equação:

$$\phi(w) = \frac{w^2}{2} \tag{6}$$

A MVS ainda possibilita encontrar um hiperplano que minimize a ocorrência de erros de classificação nos casos em que uma perfeita separação entre as duas classes não for possível. Isso graças à inclusão de variáveis de folga, que permitem que as restrições presentes na Equação 5 sejam quebradas.

O problema de otimização passa a ser então a minimização da Equação 7, de acordo com a restrição imposta na Equação 8. *C* é um parâmetro de treinamento que estabelece um equilíbrio entre a complexidade do modelo e o erro de treinamento e deve ser selecionado pelo usuário.

$$\phi(w, \xi) = \frac{w^2}{2} + C \sum_{i=1}^{N} \xi_i \tag{7}$$

sujeito à

$$y_i\big((w.x_i) + b\big) + \xi_i \geq 1 \tag{8}$$

Através da teoria dos multiplicadores de Lagrange, chega-se à Equação 9. O objetivo então passa a ser encontrar os multiplicadores de Lagrange a_i ótimos que satisfaçam a Equação 10 (CHAVES, 2006).

$$L(a) = \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i,j=1}^{N} a_i a_j \, y_i y_j \, (x_i, x_j) \tag{9}$$

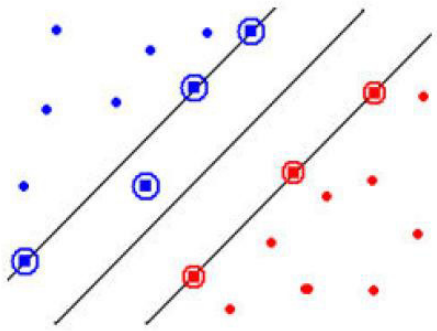
$$\sum_{i=1}^{N} a_i y_i = 0, \quad 0 \leq a_i \leq C \tag{10}$$

Apenas os pontos onde a restrição dada pela Equação 5 é exatamente igual à unidade têm correspondentes $a_i \neq 0$. Esses pontos são chamados de vetores de suporte, pois se localizam geometricamente sobre as margens. Tais pontos têm fundamental importância na

definição do hiperplano ótimo, pois os mesmos delimitam a margem do conjunto de treinamento. A Figura 8 destaca os pontos que representam os vetores de suporte.

Os pontos além da margem não influenciam decisivamente na determinação do hiperplano, enquanto que os vetores de suporte, por terem pesos não nulos, são decisivos.

Figura 8 - Vetores de Suporte (destacado por círculos).



Fonte: (NASCIMENTO, 2012).

Para que a MVS possa classificar amostras que não são linearmente separáveis, é necessária uma transformação não-linear que transforme o espaço de entrada (dados) para um novo espaço (espaço de características).

Esse espaço deve apresentar dimensão suficientemente grande e, através dele, a amostra pode ser linearmente separável. Dessa maneira, o hiperplano de separação é definido como uma função linear de vetores retirados do espaço de características em vez do espaço de entrada original. Essa construção depende do cálculo de uma função K de núcleo de um produto interno (HAYKIN, 2001). A função K pode realizar o mapeamento das amostras para um espaço de dimensão muito elevada sem aumentar a complexidade dos cálculos.

A Equação 11 mostra o resultado da Equação 9 com a utilização de um núcleo K.

$$L(a) = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i,j=1}^{N} a_i a_j \, y_i y_j K(x_i, x_j) \qquad (11)$$

Uma importante família de funções de núcleo é a função de base radial, muito utilizada em problemas de reconhecimento de padrões e também empregada neste trabalho. A função de base radial é definida por:

$$K(x_i, y_j) = \exp(-\gamma \parallel x_i - y_j \parallel^2) \tag{12}$$

onde $\gamma = 1/\sigma^2$, sendo σ a variância.

2.8 Validação dos Resultados

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

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

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

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

$$SE = \frac{VP}{VP+FN} \tag{13}$$

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

$$ES = \frac{VN}{VN+FP} \tag{14}$$

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

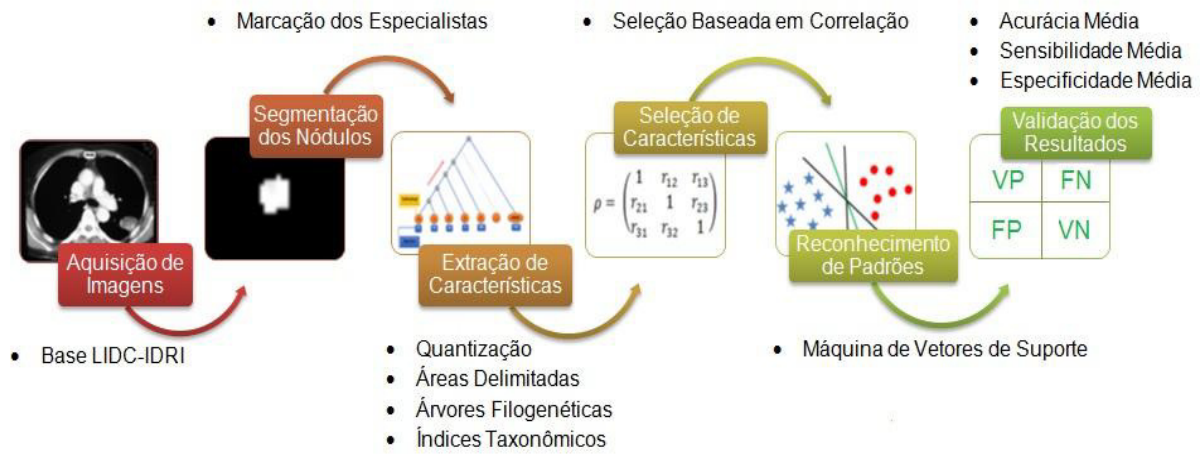
$$AC = \frac{VP+VN}{VP+FN+VN+FP} \tag{15}$$

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

3. METODOLOGIA

Neste capítulo são descritas as etapas utilizadas na metodologia proposta para a classificação de nódulos pulmonares em exames de tomografia computadorizada. A metodologia está dividida em seis grandes etapas como descrita na Figura 9. Em síntese, a primeira etapa é a aquisição das imagens que foram obtidas da base de dados de imagens de exames de tomografia computadorizada LIDC-IDRI. Na segunda etapa é realizada a segmentação dos nódulos. Na terceira, é feita a extração de características dos nódulos pulmonares, utilizando os índices taxonômicos. Na quarta etapa é realizada a seleção de características. Após essa etapa, é feita a classificação utilizando a MVS e por fim, os resultados são avaliados.

Figura 9 - Etapas da Metodologia.



3.1 Aquisição das Imagens

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

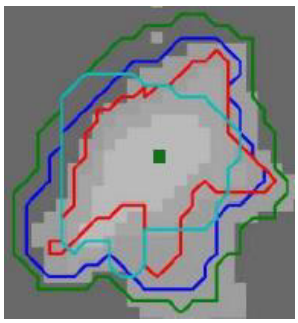
Na base LIDC-IDRI, as imagens estão no formato DICOM e possuem 16 bits por *voxel*. A base fornece um arquivo em formato XML com a informação do contorno ao longo

das fatias, além de algumas características como esfericidade, textura, malignidade e etc (a estas é indicado um valor de 1 a 5), para aqueles nódulos pulmonares maiores que 3 mm, e apenas a informação sobre o centróide para aqueles inferiores a 3 mm.

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

Não há imposição para que haja consenso, todos os nódulos indicados pela revisão dos radiologistas são apurados e gravados. Sendo assim, é possível ter diferentes diagnósticos para um mesmo nódulo. Considera-se, então, neste trabalho, apenas uma instância por nódulo, objetivando minimizar o impacto da subjetividade nos exames. No entanto, não existe nenhuma indicação na anotação dos radiologistas (arquivo XML) sobre quais informações se referem ao mesmo nódulo. Para esta tarefa, então, calcula-se o ponto central dos nódulos posteriormente verificando se as coordenadas desse ponto se encontram na região de um nódulo apurado por outro especialista. A Figura 10 ilustra o processo:

Figura 10 - Ilustração do resumo das marcações dos nódulos**.**

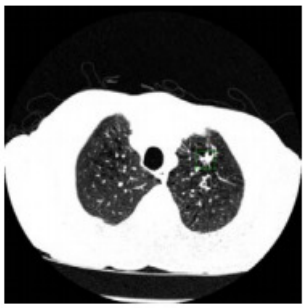


Fonte: (NASCIMENTO, 2012).

Na Figura 10, as linhas coloridas representam os contornos definidos pelos especialistas individualmente. O quadrado verde, no centro, refere-se ao centróide calculado para o contorno da mesma cor. Conforme o cálculo, as coordenadas desse centróide se encontram nas áreas delimitadas por outros especialistas. Dessa forma, considera-se, neste trabalho, que se trata do mesmo nódulo e, portanto, só deve ser aceita uma instância, referente

àquela marcação do nódulo que possuir a maior área de contorno. Após o cálculo de quais nódulos foram anotados por mais de um especialista, e de selecionar quais as instâncias correspondentes que serão utilizadas, é feito o resumo do diagnóstico quanto à malignidade ou benignidade. O diagnóstico já está presente para cada nódulo gravado na base, em uma escala de malignidade de cinco níveis representados no arquivo XML por números de 1 a 5 (“altamente improvável”, “moderadamente improvável”, “indeterminado”, “moderadamente provável”, ou “altamente provável”, respectivamente). Para o resumo, então, utilizam-se as informações conforme levantamento da etapa anterior para que seja efetuado o cálculo segundo apresentado em (JABON, RAICU & FURST, 2009), em que os valores das características pertencentes ao mesmo nódulo são reduzidos a um único valor através do cálculo da moda ou mediana. Esse processo é ilustrado na Figura 11 destacando-se com o retângulo a característica de malignidade que será utilizada no trabalho.

Figura 11 - Ilustração do resumo do diagnóstico dos nódulos.



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

Fonte: (NASCIMENTO, 2012).

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

3.2 Segmentação 3D dos Nódulos

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

3.3 Extração de Características

Após a segmentação dos nódulos, os mesmos são encaminhados à fase de extração de características de textura. Nesta fase, foram realizados experimentos com a aplicação da técnica de quantização uniforme em três níveis: 8, 12 e 16 bits. Para a descrição da textura dos objetos, foram utilizados os Índices de Diversidade Taxonômica e Distinção Taxonômica. Como esses índices são baseados na distância filogenética (contabilização do número de arestas) a partir da arquitetura de determinada árvore, foram desenvolvidas três formas de árvores para este trabalho. Os outros requisitos necessários para a geração da árvore são as espécies (unidade de Hounsfield) e os indivíduos (voxels) adquiridos com base nas duas abordagens, que foram: máscara interna e máscara externa. Essas abordagens foram utilizadas para encontrar padrões de textura que melhor descrevam os nódulos pulmonares benigno e maligno.

3.3.1 Abordagem com Máscara Interna

A proposta desta abordagem é baseada em descobrir padrões de diversidade nas áreas próximas a borda e internamente. Essas regiões foram geradas a partir de máscaras, que são imagens binárias. A primeira máscara foi criada com a binarização da região de interesse (ROI) original, a segunda com base na diminuição da escala em relação à primeira pelo centro de massa e as sucessoras máscaras foram adquiridas a partir das suas anteriores na sequência até a mais interna. O esquema do procedimento para geração das máscaras e, consequentemente, de suas áreas de interesse é apresentado na Figura 12.

Figura 12 - Procedimento da criação das cinco máscaras internas.

Neste trabalho foi definido o valor de 20% na diminuição da escala, pois nos testes foi verificado que os melhores resultados foram obtidos utilizando-se 5 máscaras de imagem com essa proporção de escalonamento, sendo que a 1ª máscara não foi escalada (imagem original binarizada).

3.3.2 Abordagem com Máscara Externa

Esta abordagem é similar à abordagem com máscara interna. As máscaras externas são formadas pela diferença entre as máscaras internas, sendo que a primeira foi criada pela diferença entre a primeira e a segunda da interna; a segunda, entre a terceira com a segunda máscara interna. As demais máscaras externas foram originadas da sequência das diferenças até o último par das máscaras internas. Na Figura 13, são demonstrados os passos para criação das máscaras externas.

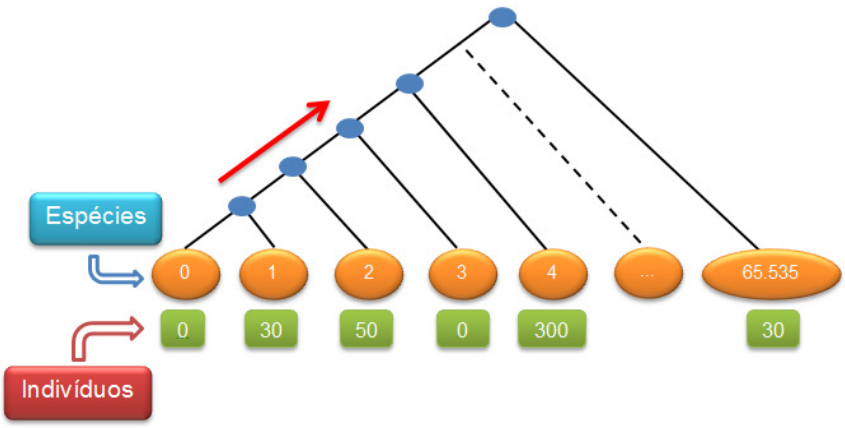
Figura 13 - Procedimento da criação das quatro máscaras externas.

3.3.3 Árvore 1 – Árvore Enraizada na Forma de Cladograma Inclinado

Após a geração das abordagens descritas nas Seções 3.3.1 e 3.3.2 são criadas as árvores para cada uma das áreas delimitadas (5 máscaras internas e 4 máscaras externas). Na Figura 14, é mostrada uma árvore, na qual as espécies são unidade Hounsfield (UH), que podem variar entre -32.768 e +32.768. Assim, foi aplicado uma mudança simples para que todos os valores ficassem positivos, com objetivo apenas de tornar mais simples o cálculo dos índices. Esta mudança ocorre quando é deslocado o menor valor negativo para que se inicie de zero, assim torna-se possível identificar até 65.536 espécies.

A relação entre as espécies é feita da esquerda para direita, conforme indicado pela seta vermelha na Figura 14. Dessa forma, a primeira relação é entre as espécies 0 e 1, que possui duas arestas ligando as mesmas (ω_{0-1}), como mostra a Figura 15 (a), e também é efetuado o cálculo do denominador das Equações 2 e 3. Na segunda relação, são três arestas (ω_{0-2}) que ligam as espécies 0 e 2 Figura 15 (b) . Dessa forma, o último relacionamento da espécie 0 é com a espécie 65.535, que possui 65.536 arestas ($\omega_{0-65.535}$).

Figura 14 - Árvore 1: Árvore enraizada na forma de cladograma inclinado.



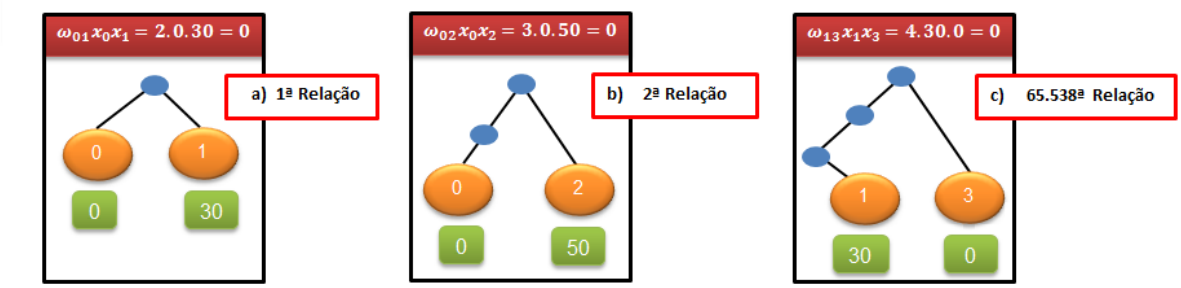
A próxima espécie a se relacionar com as demais é a um 1, em que a primeira relação tem início a partir da espécie 2 (ω_{1-2} = três arestas), como mostra a condição $i < j$ (Equação 2 e 3), ou seja, é feita a combinação de uma espécie com outras que ainda não tenham sido processadas, portanto, (ω_{1-0} é o mesmo que ω_{0-1}). Em seguida, a espécie 1 é combinada com a espécie 3 (ω_{1-3} = quatro arestas), como mostra na Figura 15 (c). O último

relacionamento da espécie 1 é feito com a espécie 65.535 que possui 65.536 arestas ($\omega_{1-65.536}$). O restante das combinações segue a mesma regra de não fazer relação com espécies que já foram combinadas.

3.3.4 Árvore 2 – Árvore Enraizada na Forma de Cladograma Inclinado Excluindo as Espécies sem Indivíduos

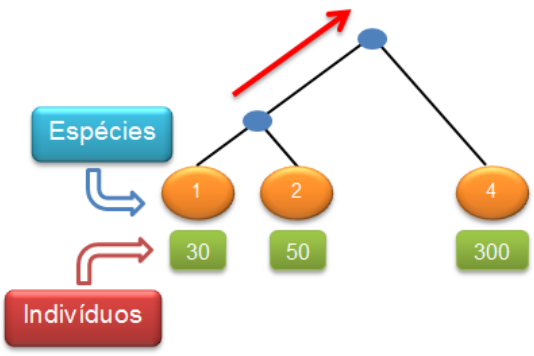
Fundamentado na mesma lógica do cálculo dos índices com base na árvore anterior, foi desenvolvida outra arquitetura de árvore que tem como destaque a eliminação das espécies que não possuem indivíduos, resultando consequentemente na reorganização das arestas para as espécies remanescentes.

Figura 15 - Descrição da quantidade de arestas das espécies 0 com 1 (a), 0 com 2 (b) e 1 com 3 (c).



Supondo que o fragmento retirado da árvore mostrado na Figura 14 tenha somente indivíduos nas espécies 1, 2 e 4, o novo modelo de árvore possui somente essas espécies e segue a mesma arquitetura descrita anteriormente, cujo resultado é apresentado na Figura 16.

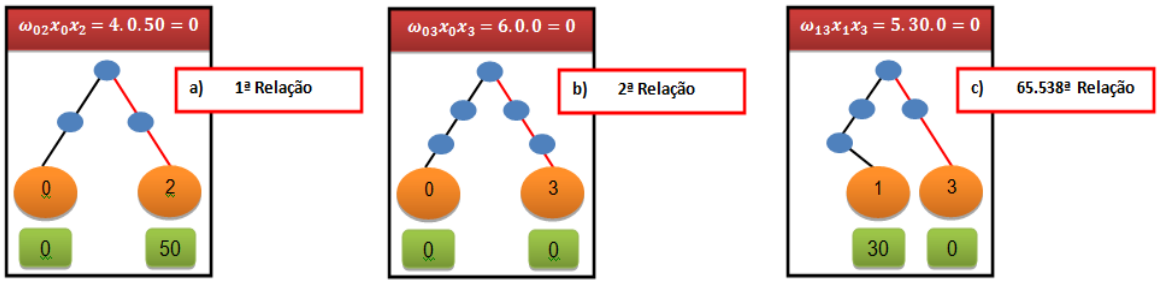
Figura 16 - Árvore 2: Modelo criado a partir da Árvore 1, com eliminação das espécies sem indivíduos.



3.3.5 Árvore 3 – Árvore Enraizada na Forma de Cladograma Inclinado Modificando as Arestas

Já a terceira árvore proposta tem o mesmo processo de combinação entre as espécies da Árvore 1, sendo que a única diferença é na quantidade de arestas do lado direito do nó ancestral entre as espécies mostrado na Figura 14. Assim, a Figura 17 descreve o mesmo procedimento da Figura 14, sendo destacado em vermelho as arestas que eram somente uma na Árvore 1 e a primeira combinação de uma espécie com as outras que não tem mudança (segue mesmo exemplo da Figura 15).

Figura 17 - Descrição da quantidade de arestas das espécies 0 com 2 (a), 0 com 3 (b) e 1 com 3 (c).



3.4 Seleção de Características

Para realizar a seleção de características que melhor discriminam as classes maligno e benigno, foi utilizada a técnica de seleção de características baseada em correlação com o *best-first-search*, que é fornecida no software Weka (HALL *et al*., 2009).

3.5 Reconhecimento de Padrões

O processo de classificação objetiva analisar os padrões obtidos através das extrações de características propostas usando o classificador MVS.

Para efeito de validação dos resultados as amostras foram divididas, aleatoriamente, em dois grupos: grupo de treino e grupo de teste.

Durante a etapa de treinamento é gerado o modelo com os vetores de suporte utilizados pela MVS na etapa de teste. A etapa de treinamento desconhece por completo as amostras de testes. Esse mecanismo pretende se assemelhar com condições reais.

Para a MVS foi utilizado o núcleo função de base radial. Dessa maneira é necessário estimar o melhor peso C e também o valor γ para a função radial, além disso, foram atribuídos pesos para as classes devido ao desbalanceamento entre suas amostras, peso 1 para à classe benigno e peso 3 para à classe maligno. Os valores de C e γ são estimados com a base de treino através de uma busca exaustiva realizado pelo script desenvolvido em PYTHON: *grid.py* presente no pacote LIBSVM (CHANG & LIN,2003).

3.6 Validação dos Resultados

Após a finalização da etapa de reconhecimento de padrões, é necessário validar os resultados e discutir prováveis melhorias. Essa metodologia usa métricas comumente empregadas em sistemas CAD/CADx, e aceitas pela sociedade para análise de desempenho de sistemas baseados em processamento de imagens. Estas métricas são sensibilidade, especificidade e acurácia.

Tais métricas têm o objetivo de medir o desempenho da metodologia como satisfatória ou não, além de ajudar a identificar pontos positivos e negativos para melhoria futura deste trabalho na fase de treinamento e teste.

4. RESULTADOS E DISCUSSÃO

Nesse capítulo serão apresentados e discutidos os resultados obtidos com a utilização da metodologia proposta no trabalho para classificação de nódulos pulmonares de tomografia computadorizada em maligno e benigno. Mas, antes disso, foram realizados testes trabalhando com as três classes (benigno, maligno e indeterminado) que apresentaram resultados insatisfatórios, pois a classe indeterminada não possui um padrão que a diferencie das demais classes, porque um nódulo diagnosticado como indeterminado será, após um tempo de avaliação, diagnosticado como maligno ou benigno, ou seja, irá pertencer a uma das duas classes.

Para a realização dos testes, a partir da etapa de extração e seleção de características, a base de amostras foi organizada em dois grupos para serem utilizados no classificador MVS: grupo de treino e grupo de teste, com proporções de 20% e 80%, 40% e 60%, 60% e 40%, 80% e 20%, respectivamente. Todos os valores das bases que estavam no conjunto $\mathbb{R}^+$ (conjunto de números reais não-negativos) foram normalizados entre -1 a 1 para ajudar o classificador a convergir com maior facilidade na etapa de treinamento.

Como resultado da obtenção das marcações dos especialistas e da etapa de segmentação foram gerados 2.393 nódulos, sendo 1.011 benignos, 394 malignos e 988 indeterminados (Seção 3.2). Em seguida, os nódulos foram submetidos à técnica de quantização uniforme em três níveis: 8, 12 e 16 bits (Seção 2.3.1). Após a quantização, cada nódulo quantizado gerou 9 áreas delimitadas referentes as máscaras internas e máscaras externas (Seção 3.3.1 e Seção 3.3.2), e por último, foram aplicados os cálculos dos índices taxonômicos (Seção 2.5.2) para cada modelo de árvore filogenética (Seção 3.3.3, Seção 3.3.4 e Seção 3.3.5), gerando um total de 54 características para cada nódulo pulmonar.

Os testes foram divididos em duas seções (Seção 4.1 e Seção 4.2). A Seção 4.1 apresenta os resultados obtidos com as três classes (benigno, maligno e indeterminado) e a Seção 4.2 apresenta os resultados com apenas as classes benigno e maligno, além dos resultados obtidos com a aplicação da técnica de seleção de características baseada em correlação (Seção 2.6.1) para cada modelo de árvore filogenética e, por último, um teste com todas as árvores juntas.

4.1 Testes com as Classes Benigno, Maligno e Indeterminado

Nessa seção serão apresentados os resultados obtidos com as classes benigno, maligno e indeterminado, totalizando 2.393 amostras de nódulos pulmonares, sendo 1.011 benignos, 394 malignos e 988 indeterminados. Devido ao desbalanceamento das amostras foram atribuídos pesos às classes, peso 1 para à classe benigno, peso 3 para à classe maligno e peso 1 para à classe indeterminado, obedecendo às proporções estabelecidas no Capítulo 4.

Os resultados obtidos para cada árvore, além de um teste com todas as árvores juntas, são apresentados na forma de matriz de confusão que é criada classificando-se todos os casos do modelo em categorias (classes), determinando se o valor previsto (dado pelo classificador) correspondeu ao valor real. As linhas na matriz representam os valores reais para o modelo, sendo que as colunas representam os valores previstos. Todos os casos em cada categoria são contabilizados e os totais são exibidos na matriz.

4.1.1 Árvore 1

Na Tabela 1 são apresentados os resultados obtidos pelo modelo da árvore 1 (Seção 3.3.3), em suas quatro proporções de treino e teste, sem a aplicação da técnica de seleção de características baseada em correlação.

Tabela 1 - Resultados obtidos pela Árvore 1 para as 3 Classes.

Árvore 1			
Proporção 20% Treino e 80% Teste			
	B	M	I
B	214	159	425
M	17	227	72
I	88	258	454

Proporção 40% Treino e 60% Teste			
	B	M	I
B	143	76	380
M	5	172	58
I	27	174	401

Proporção 60% Treino e 40% Teste			
	B	M	I
B	101	36	247
M	5	119	48
I	19	108	274

Proporção 80% Treino e 20% Teste			
	B	M	I
B	56	18	128
M	3	56	18
I	10	56	134

B = Classe Benigno M = Classe Maligno I = Classe Indeterminado

Na proporção 20% para treino e 80% para teste, a árvore 1 sem a seleção de características obteve uma taxa de acurácia de 46,80%, a proporção 40% para treino e 60% para teste obteve uma taxa de acurácia de 49,90%, já a proporção 60% para treino e 40% para

teste obteve uma taxa de acurácia de 51,60% e na proporção 80% para treino e 20% para teste obteve uma taxa de acurácia de 51,40%.

Na Tabela 2 são apresentados os resultados obtidos pelo modelo da árvore 1, em suas quatro proporções de treino e teste, com a aplicação da técnica de seleção de características baseada em correlação.

Tabela 2 - Resultados obtidos pela Árvore 1 para as 3 Classes com Seleção de Características.

Árvore 1 - Com Seleção de Características							
Proporção 20% Treino e 80% Teste				Proporção 40% Treino e 60% Teste			
	B	M	I		B	M	I
B	333	154	311	B	161	88	350
M	32	234	50	M	11	166	58
I	235	255	310	I	58	178	366
Proporção 60% Treino e 40% Teste				Proporção 80% Treino e 20% Teste			
	B	M	I		B	M	I
B	109	44	231	B	66	25	111
M	11	120	41	M	5	54	18
I	33	116	252	I	14	61	125
B = Classe Benigno M = Classe Maligno I = Classe Indeterminado							

Na proporção 20% para treino e 80% para teste, a árvore 1 com a seleção de características obteve uma taxa de acurácia de 45,80%, a proporção 40% para treino e 60% para teste obteve uma taxa de acurácia de 48,30%, já a proporção 60% para treino e 40% para teste obteve uma taxa de acurácia de 50,30% e na proporção 80% para treino e 20% para teste obteve uma taxa de acurácia de 51,10%.

Analisando as Tabela 1 e Tabela 2, verifica-se que a aplicação da técnica de seleção de características não melhorou os resultados obtidos em nenhuma das quatro proporções estabelecidas.

4.1.2 Árvore 2

Na Tabela 3 são apresentados os resultados obtidos pelo modelo da árvore 2 (Seção 3.3.4), em suas quatro proporções de treino e teste, sem a aplicação da técnica de seleção de características baseada em correlação.

Tabela 3 - Resultados obtidos pela Árvore 2 para as 3 Classes.

Árvore 2			
Proporção 20% Treino e 80% Teste			
	B	M	I
B	532	38	228
M	41	223	52
I	395	121	284

Proporção 40% Treino e 60% Teste			
	B	M	I
B	276	37	286
M	18	172	45
I	162	126	314

Proporção 60% Treino e 40% Teste			
	B	M	I
B	242	19	123
M	15	123	34
I	156	82	163

Proporção 80% Treino e 20% Teste			
	B	M	I
B	128	11	63
M	8	54	15
I	72	44	84

B = Classe Benigno M = Classe Maligno I = Classe Indeterminado

Na proporção 20% para treino e 80% para teste, a árvore 2 sem a seleção de características obteve uma taxa de acurácia de 54,30%, a proporção 40% para treino e 60% para teste obteve uma taxa de acurácia de 53,10%, já a proporção 60% para treino e 40% para teste obteve uma taxa de acurácia de 55,20% e na proporção 80% para treino e 20% para teste obteve uma taxa de acurácia de 55,50%.

Na Tabela 4 são apresentados os resultados obtidos pelo modelo da árvore 2, em suas quatro proporções de treino e teste, com a aplicação da técnica de seleção de características baseada em correlação.

Tabela 4 - Resultados obtidos pela Árvore 2 para as 3 Classes com Seleção de Características.

Árvore 2 - Com Seleção de Características			
Proporção 20% Treino e 80% Teste			
	B	M	I
B	469	36	293
M	45	218	53
I	359	128	313

Proporção 40% Treino e 60% Teste			
	B	M	I
B	254	31	314
M	20	172	43
I	164	115	323

Proporção 60% Treino e 40% Teste			
	B	M	I
B	246	15	123
M	18	124	30
I	166	77	158

Proporção 80% Treino e 20% Teste			
	B	M	I
B	118	10	74
M	12	55	10
I	72	44	84

B = Classe Benigno M = Classe Maligno I = Classe Indeterminado

Na proporção 20% para treino e 80% para teste, a árvore 2 com a seleção de características obteve uma taxa de acurácia de 52,20%, a proporção 40% para treino e 60% para teste obteve uma taxa de acurácia de 52,20%, já a proporção 60% para treino e 40% para teste obteve uma taxa de acurácia de 55,20% e na proporção 80% para treino e 20% para teste obteve uma taxa de acurácia de 53,70%.

Analisando as Tabela 3 e Tabela 4, verifica-se que a aplicação da técnica de seleção de características não melhorou os resultados obtidos em nenhuma das quatro proporções estabelecidas.

4.1.3 Árvore 3

Na Tabela 5 são apresentados os resultados obtidos pelo modelo da árvore 3 (Seção 3.3.5), em suas quatro proporções de treino e teste, sem a aplicação da técnica de seleção de características baseada em correlação.

Tabela 5 - Resultados obtidos pela Árvore 3 para as 3 Classes.

Árvore 3

Proporção 20% Treino e 80% Teste

	B	M	I
B	193	192	413
M	14	232	70
I	62	313	425

Proporção 40% Treino e 60% Teste

	B	M	I
B	140	86	373
M	4	163	68
I	24	173	405

Proporção 60% Treino e 40% Teste

	B	M	I
B	91	37	256
M	6	116	50
I	11	105	285

Proporção 80% Treino e 20% Teste

	B	M	I
B	47	19	136
M	1	55	21
I	6	56	138

B = Classe Benigno M = Classe Maligno I = Classe Indeterminado

Na proporção 20% para treino e 80% para teste, a árvore 3 sem a seleção de características obteve uma taxa de acurácia de 44,40%, a proporção 40% para treino e 60% para teste obteve uma taxa de acurácia de 49,30%, já a proporção 60% para treino e 40% para teste obteve uma taxa de acurácia de 51,40% e na proporção 80% para treino e 20% para teste obteve uma taxa de acurácia de 50,10%.

Na Tabela 6 são apresentados os resultados obtidos pelo modelo da árvore 3, em suas quatro proporções de treino e teste, com a aplicação da técnica de seleção de características baseada em correlação.

Tabela 6 - Resultados obtidos pela Árvore 3 para as 3 Classes com Seleção de Características.

Árvore 3 - Com Seleção de Características							
Proporção 20% Treino e 80% Teste				Proporção 40% Treino e 60% Teste			
	B	M	I		B	M	I
B	248	122	428	B	150	103	346
M	35	215	66	M	8	169	58
I	134	220	446	I	47	182	373
Proporção 60% Treino e 40% Teste				Proporção 80% Treino e 20% Teste			
	B	M	I		B	M	I
B	99	58	227	B	47	31	124
M	5	125	42	M	2	54	21
I	22	125	254	I	3	67	130
B = Classe Benigno M = Classe Maligno I = Classe Indeterminado							

Na proporção 20% para treino e 80% para teste, a árvore 3 com a seleção de características obteve uma taxa de acurácia de 47,50%, a proporção 40% para treino e 60% para teste obteve uma taxa de acurácia de 48,20%, já a proporção 60% para treino e 40% para teste obteve uma taxa de acurácia de 49,90% e na proporção 80% para treino e 20% para teste obteve uma taxa de acurácia de 47,90%.

Analisando as Tabela 5 e Tabela 6, verifica-se que a aplicação da técnica de seleção de características obteve taxa de acurácia maior apenas na proporção 20% para treino e 80% para teste.

4.1.4 Todas as Árvores Juntas

Na Tabela 7 são apresentados os resultados obtidos pela junção dos três modelos de árvores filogenéticas, em suas quatro proporções de treino e teste, sem a aplicação da técnica de seleção de características baseada em correlação.

Tabela 7 - Resultados obtidos pela Junção de todas as Árvores para as 3 Classes.

Todas as Árvores Juntas							
Proporção 20% Treino e 80% Teste				Proporção 40% Treino e 60% Teste			
	B	M	I		B	M	I
B	247	40	511	B	149	28	422
M	20	224	72	M	5	171	59
I	91	135	574	I	36	125	441
Proporção 60% Treino e 40% Teste				Proporção 80% Treino e 20% Teste			
	B	M	I		B	M	I
B	103	17	264	B	56	7	139
M	7	123	42	M	1	54	22
I	28	83	290	I	7	40	153
B = Classe Benigno M = Classe Maligno I = Classe Indeterminado							

Na proporção 20% para treino e 80% para teste, a junção de todas as árvores sem a seleção de características obteve uma taxa de acurácia de 54,60%, a proporção 40% para treino e 60% para teste obteve uma taxa de acurácia de 53,00%, já a proporção 60% para treino e 40% para teste obteve uma taxa de acurácia de 53,90% e na proporção 80% para treino e 20% para teste obteve uma taxa de acurácia de 54,90%.

Na Tabela 8 são apresentados os resultados obtidos pela junção dos três modelos de árvores filogenéticas, em suas quatro proporções de treino e teste, com a aplicação da técnica de seleção de características baseada em correlação.

Tabela 8 - Resultados obtidos pela Junção de todas as Árvores para as 3 Classes com Seleção de Características.

Todas as Árvores Juntas - Com Seleção de Características			
Proporção 20% Treino e 80% Teste			
	B	M	I
B	356	34	408
M	23	213	80
I	234	115	451

Proporção 40% Treino e 60% Teste			
	B	M	I
B	185	32	382
M	12	168	55
I	72	118	412

Proporção 60% Treino e 40% Teste			
	B	M	I
B	139	16	229
M	8	123	41
I	51	74	276

Proporção 80% Treino e 20% Teste			
	B	M	I
B	71	11	120
M	5	54	18
I	20	41	139

B = Classe Benigno M = Classe Maligno I = Classe Indeterminado

Na proporção 20% para treino e 80% para teste, a junção de todas as árvores com a seleção de características obteve uma taxa de acurácia de 53,30%, a proporção 40% para treino e 60% para teste obteve uma taxa de acurácia de 53,30%, já a proporção 60% para treino e 40% para teste obteve uma taxa de acurácia de 56,20% e na proporção 80% para treino e 20% para teste obteve uma taxa de acurácia de 55,10%.

Analisando as Tabela 7 e Tabela 8, verifica-se que a aplicação da técnica de seleção de características obteve taxas de acurácias maiores em todas as proporções, exceto a proporção 20% para treino e 80% para teste.

Sintetizando os resultados obtidos com as três classes (benigno, maligno e indeterminado), a melhor acurácia alcançada foi obtida na junção de todas as árvores com a aplicação da seleção de características na proporção 60% para treino e 40% para teste. Já o

pior resultado, foi obtido pelo modelo da árvore 3 na proporção 20% para treino e 80% para teste sem a aplicação da seleção de características.

4.2 Testes com as Classes Benigno e Maligno

Nessa seção serão apresentados os resultados obtidos com as classes benigno e maligno, totalizando 1.405 amostras de nódulos pulmonares, sendo 1.011 benignos e 394 malignos. Devido ao desbalanceamento das amostras foram atribuídos pesos às classes, peso 1 para à classe benigno e peso 3 para à classe maligno, obedecendo às proporções estabelecidas no Capítulo 4. Para cada proporção, os testes foram repetidos aleatoriamente 5 vezes com o intuito de verificar se os acertos em todas as repetições apresentam semelhança dos valores altos e pequena diferença entre eles, assim representando que o padrão de textura discrimina bem as amostras de nódulos benignos e malignos.

As medidas de validação dos resultados utilizadas são: sensibilidade média, especificidade média e acurácia média, além do desvio padrão. É utilizada ainda uma nomenclatura referente a cada quantização, sendo Q8, Q12 e Q16 concernentes a utilização das imagens com quantização em 8, 12 e 16 bits respectivamente, além das siglas MI1, MI2, MI3, MI4 e MI5 para as máscaras internas e ME1, ME2, ME3 e ME4 para as máscaras externas.

4.2.1 Árvore 1

Na Tabela 9 são apresentados os resultados obtidos pelo modelo da árvore 1, em suas quatro proporções de treino e teste, sem a aplicação da técnica de seleção de características baseada em correlação, com 54 características.

Conforme a Tabela 9, o melhor resultado obtido foi de 84,65% de acurácia média, 80,84% de sensibilidade média e 86,15% de especificidade média na proporção 40% para treino e 60% para teste. O pior resultado encontrado foi na proporção 20% para treino e 80% para teste com acurácia média de 82,80%, sensibilidade média de 74,80% e 86,03% de especificidade média.

Tabela 9 - Resultados obtidos pela Árvore 1 para as 2 Classes.

Árvore 1									
Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade	Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade
20%-80%	Teste 1	82,30%	77,88%	84,06%	40%-60%	Teste 1	86,12%	77,18%	89,70%
	Teste 2	82,21%	65,02%	89,14%		Teste 2	83,51%	78,79%	85,29%
	Teste 3	81,85%	76,90%	83,90%		Teste 3	84,82%	82,23%	85,86%
	Teste 4	85,41%	73,15%	90,38%		Teste 4	84,70%	80,17%	86,47%
	Teste 5	82,21%	81,03%	82,66%		Teste 5	84,10%	85,84%	83,44%
	Média	82,80%	74,80%	86,03%		Média	84,65%	80,84%	86,15%
	Desvio Padrão	1,32	5,50	3,11		Desvio Padrão	0,87	3,00	2,04
60%-40%	Teste 1	84,52%	83,85%	84,79%	80%-20%	Teste 1	80,43%	86,57%	78,50%
	Teste 2	83,45%	81,01%	84,41%		Teste 2	83,63%	73,42%	87,62%
	Teste 3	84,70%	80,12%	86,53%		Teste 3	86,12%	87,32%	85,71%
	Teste 4	85,23%	83,23%	86,00%		Teste 4	80,43%	77,78%	81,68%
	Teste 5	84,88%	77,42%	87,71%		Teste 5	84,34%	85,71%	83,82%
	Média	84,56%	81,13%	85,89%		Média	82,99%	82,16%	83,47%
	Desvio Padrão	0,60	2,31	1,19		Desvio Padrão	2,24	5,55	3,17

Analisando a Tabela 10, foi verificado que a diferença entre o melhor e o pior resultado é de apenas 1,85% de acurácia média, possuindo um valor abaixo de 1 no desvio padrão na métrica da acurácia média.

Tabela 10 - Resultados das médias das proporções obtidas pela Árvore 1.

Árvore 1			
Proporção	Acurácia	Sensibilidade	Especificidade
Média 20%-80%	82,80%	74,80%	86,03%
Média 40%-60%	84,65%	80,84%	86,15%
Média 60%-40%	84,56%	81,13%	85,89%
Média 80%-20%	82,99%	82,16%	83,47%
Desvio Padrão	0,86	2,89	1,11

Na Tabela 11 são apresentados os resultados obtidos pelo modelo da árvore 1 em suas quatro proporções de treino e teste, com a aplicação da técnica de seleção de características baseada em correlação, com 8 características selecionadas (Q8/MI1/Δ*, Q8/MI2/Δ*, Q8/ME2/Δ, Q12/ME4/Δ, Q12/MI5/Δ, Q12/MI5/Δ*, Q16/MI5/Δ, e Q16/MI5/Δ*).

Conforme a Tabela 11, o melhor resultado obtido foi de 83,92% de acurácia média, 75,14% de sensibilidade média e 87,44% de especificidade média na proporção 80% para treino e 20% para teste. O pior resultado encontrado foi na proporção 20% para treino e 80% para teste com acurácia média de 80,50%, sensibilidade média de 72,20% e 83,74% de especificidade média.

Tabela 11 - Resultados obtidos pela Árvore 1 para as 2 Classes com Seleção de Características.

Árvore 1 - Com Seleção de Características									
Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade	Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade
20%-80%	Teste 1	81,30%	62,40%	88,90%	40%-60%	Teste 1	82,30%	72,00%	86,50%
	Teste 2	77,10%	76,70%	77,30%		Teste 2	81,50%	75,80%	83,80%
	Teste 3	78,60%	76,00%	79,60%		Teste 3	80,90%	76,30%	82,50%
	Teste 4	82,30%	72,80%	86,00%		Teste 4	81,10%	73,30%	84,20%
	Teste 5	83,20%	73,10%	86,90%		Teste 5	82,60%	79,20%	83,80%
	Média	80,50%	72,20%	83,74%		Média	81,68%	75,32%	84,16%
	Desvio Padrão	2,30	5,14	4,48		Desvio Padrão	0,66	2,50	1,30
60%-40%	Teste 1	82,90%	76,60%	85,30%	80%-20%	Teste 1	81,10%	70,70%	85,40%
	Teste 2	80,10%	74,80%	82,10%		Teste 2	85,10%	77,40%	88,30%
	Teste 3	82,20%	76,20%	84,20%		Teste 3	82,20%	75,00%	85,10%
	Teste 4	81,00%	70,60%	84,80%		Teste 4	86,10%	75,30%	90,50%
	Teste 5	82,40%	77,10%	84,20%		Teste 5	85,10%	77,30%	87,90%
	Média	81,72%	75,06%	84,12%		Média	83,92%	75,14%	87,44%
	Desvio Padrão	1,02	2,36	1,09		Desvio Padrão	1,92	2,43	2,00

Analisando a Tabela 12, foi verificado que a diferença entre o melhor e o pior resultado é de 3,42% de acurácia média, possuindo um valor acima de 1 nos desvios padrões em todas as métricas de validação.

Tabela 12 - Resultados das médias das proporções obtidas pela Árvore 1 com Seleção de Características.

Árvore 1 - Com Seleção de Características			
Proporção	Acurácia	Sensibilidade	Especificidade
Média 20%-80%	80,50%	72,20%	83,74%
Média 40%-60%	81,68%	75,32%	84,16%
Média 60%-40%	81,72%	75,06%	84,12%
Média 80%-20%	83,92%	75,14%	87,44%
Desvio Padrão	1,24	1,29	1,50

Analisando as Tabela 10 e Tabela 12, verifica-se que a aplicação da técnica de seleção de características obteve resultado de acurácia média maior apenas na proporção 80% para treino e 20% para teste.

4.2.2 Árvore 2

Os dados apresentados na Tabela 13 mostram os resultados obtidos pela árvore 2 com as médias das acurácias, sensibilidades e especificidades dos cinco testes realizados para cada proporção.

Tabela 13 - Resultados obtidos pela Árvore 2 para as 2 Classes.

Árvore 2									
Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade	Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade
20%-80%	Teste 1	86,48%	80,19%	88,90%	40%-60%	Teste 1	87,43%	83,20%	89,15%
	Teste 2	83,63%	77,99%	85,86%		Teste 2	85,88%	78,30%	88,82%
	Teste 3	86,48%	84,11%	87,42%		Teste 3	88,02%	78,40%	92,07%
	Teste 4	85,85%	77,04%	89,33%		Teste 4	85,29%	85,84%	85,08%
	Teste 5	86,92%	79,62%	89,81%		Teste 5	83,04%	80,87%	83,85%
	Média	85,87%	79,79%	88,26%		Média	85,93%	81,32%	87,79%
	Desvio Padrão	1,17	2,44	1,44		Desvio Padrão	1,75	2,89	2,97
60%-40%	Teste 1	89,50%	84,87%	91,22%	80%-20%	Teste 1	87,90%	85,53%	88,78%
	Teste 2	87,01%	81,82%	89,17%		Teste 2	87,54%	79,52%	90,91%
	Teste 3	87,72%	82,43%	89,61%		Teste 3	92,53%	88,57%	93,84%
	Teste 4	87,37%	86,42%	87,75%		Teste 4	86,83%	85,00%	87,56%
	Teste 5	88,43%	84,42%	89,95%		Teste 5	88,26%	89,74%	87,68%
	Média	88,01%	83,99%	89,54%		Média	88,61%	85,67%	89,75%
	Desvio Padrão	0,88	1,67	1,13		Desvio Padrão	2,02	3,56	2,37

Na árvore 2, a melhor acurácia média foi de 88,61%, sensibilidade média de 85,67% e especificidade média de 89,75% na proporção 80% para treino e 20% para teste. O pior resultado médio obtido por essa árvore foi na proporção 20% para treino e 80% para teste com acurácia média de 85,87%, sensibilidade média de 79,79% e especificidade média de 88,26%.

Analisando a Tabela 14, foi verificado que a diferença entre o melhor e o pior resultado é de 2,74% de acurácia média, possuindo um valor abaixo de 1 no desvio padrão na métrica da especificidade média.

Tabela 14 - Resultados das médias das proporções obtidas pela Árvore 2.

Árvore 2			
Proporção	Acurácia	Sensibilidade	Especificidade
Média 20%-80%	85,87%	79,79%	88,26%
Média 40%-60%	85,93%	81,32%	87,79%
Média 60%-40%	88,01%	83,99%	89,54%
Média 80%-20%	88,61%	85,67%	89,75%
Desvio Padrão	1,22	2,28	0,83

Na Tabela 15 são apresentados os resultados obtidos pelo modelo da árvore 2 em suas quatro proporções de treino e teste, com a aplicação da técnica de seleção de características baseada em correlação, com 14 características selecionadas (Q8/MI1/Δ, Q8/MI1/Δ*, Q8/MI2/Δ, Q12/MI5/Δ*, Q16/MI1/Δ*, Q16/ME1/Δ, Q16/ME1/Δ*, Q16/MI2/Δ, Q16/MI2/Δ*, Q16/ME3/Δ*, Q16/MI4/Δ, Q16/ME4/Δ*, Q16/MI5/Δ e Q16/MI5/Δ*).

Tabela 15 - Resultados obtidos pela Árvore 2 para as 2 Classes com Seleção de Características.

Árvore 2 - Com Seleção de Caracteristicas									
Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade	Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade
20%-80%	Teste 1	88,30%	75,20%	93,60%	40%-60%	Teste 1	89,10%	79,80%	92,80%
	Teste 2	89,10%	81,20%	92,10%		Teste 2	87,90%	80,30%	91,00%
	Teste 3	88,50%	81,00%	91,30%		Teste 3	88,50%	83,30%	90,30%
	Teste 4	87,80%	79,20%	91,10%		Teste 4	87,50%	78,80%	90,90%
	Teste 5	87,70%	83,90%	89,10%		Teste 5	86,80%	85,80%	87,20%
	Média	88,28%	80,10%	91,44%		Média	87,96%	81,60%	90,44%
	Desvio Padrão	0,51	2,87	1,46		Desvio Padrão	0,79	2,58	1,82
60%-40%	Teste 1	89,90%	83,80%	92,20%	80%-20%	Teste 1	87,20%	77,80%	90,40%
	Teste 2	87,70%	79,20%	91,10%		Teste 2	87,50%	81,00%	90,40%
	Teste 3	88,30%	83,90%	89,70%		Teste 3	90,40%	88,80%	91,00%
	Teste 4	87,40%	74,50%	92,20%		Teste 4	89,00%	81,50%	92,00%
	Teste 5	86,10%	84,70%	86,60%		Teste 5	86,50%	84,00%	87,40%
	Média	87,88%	81,22%	90,36%		Média	88,12%	82,62%	90,24%
	Desvio Padrão	1,24	3,88	2,09		Desvio Padrão	1,40	3,67	1,54

Conforme a Tabela 15, o melhor resultado obtido foi de 88,28% de acurácia média, 80,10% de sensibilidade média e 91,44% de especificidade média na proporção 20% para treino e 80% para teste. O pior resultado encontrado foi na proporção 60% para treino e 40% para teste com acurácia média de 87,88%, sensibilidade média de 81,22% e 90,36% de especificidade média.

Analisando a Tabela 16, foi verificado que a diferença entre o melhor e o pior resultado é de apenas 0,40% de acurácia média, possuindo um valor abaixo de 1 nos desvios padrões em todas as métricas de validação.

Tabela 16 - Resultados das médias das proporções obtidas pela Árvore 2 com Seleção de Características.

Árvore 2 - Com Seleção de Caracteristicas			
Proporção	Acurácia	Sensibilidade	Especificidade
Média 20%-80%	88,28%	80,10%	91,44%
Média 40%-60%	87,96%	81,60%	90,44%
Média 60%-40%	87,88%	81,22%	90,36%
Média 80%-20%	88,12%	82,62%	90,24%
Desvio Padrão	0,15	0,90	0,48

Analisando as Tabela 14 e Tabela 16, verifica-se que a aplicação da técnica de seleção de características obteve resultados de acurácias médias maiores na proporção 20% para treino e 80% para teste e na proporção 40% para treino e 60% para teste.

4.2.3 Árvore 3

A seguir são apresentados os resultados para as médias das acurácias, sensibilidades e especificidades dos cinco testes realizados para cada proporção na árvore 3.

Tabela 17 - Resultados obtidos pela Árvore 3 para as 2 Classes.

Árvore 3									
Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade	Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade
20%-80%	Teste 1	83,81%	74,45%	87,55%	40%-60%	Teste 1	81,85%	81,65%	81,92%
	Teste 2	81,85%	83,06%	81,40%		Teste 2	84,93%	80,77%	86,54%
	Teste 3	81,94%	78,25%	83,33%		Teste 3	81,26%	78,86%	82,24%
	Teste 4	83,19%	69,42%	88,83%		Teste 4	83,87%	79,01%	85,83%
	Teste 5	82,65%	83,71%	82,24%		Teste 5	83,16%	78,24%	85,10%
	Média	82,69%	77,78%	84,67%		Média	83,01%	79,71%	84,33%
	Desvio Padrão	0,74	5,37	2,97		Desvio Padrão	1,33	1,29	1,89
60%-40%	Teste 1	84,34%	81,46%	85,40%	80%-20%	Teste 1	81,85%	76,83%	83,92%
	Teste 2	85,05%	83,23%	85,75%		Teste 2	84,34%	86,90%	83,25%
	Teste 3	82,92%	82,94%	82,91%		Teste 3	85,05%	81,94%	86,12%
	Teste 4	82,03%	78,85%	83,25%		Teste 4	82,56%	85,71%	81,37%
	Teste 5	83,99%	82,17%	84,69%		Teste 5	85,41%	87,01%	84,80%
	Média	83,67%	81,73%	84,40%		Média	83,84%	83,68%	83,89%
	Desvio Padrão	1,07	1,57	1,14		Desvio Padrão	1,40	3,89	1,59

Já a árvore 3 (Tabela 17) apresenta sua melhor acurácia média de 83,84% na proporção de 80% para treino e 20% para teste com 83,68% e 83,89% de sensibilidade e especificidade média respectivamente. E no pior resultado, acurácia média de 82,69%, sensibilidade média de 77,78% e especificidade média de 84,67%.

Analisando a Tabela 18, foi verificado que a diferença entre o melhor e o pior resultado é de apenas 1,15% de acurácia média, possuindo um valor acima de 1 no desvio padrão na métrica da sensibilidade média.

Tabela 18 - Resultados das médias das proporções obtidas pela Árvore 3.

Árvore 3			
Proporção	Acurácia	Sensibilidade	Especificidade
Média 20%-80%	82,69%	77,78%	84,67%
Média 40%-60%	83,01%	79,71%	84,33%
Média 60%-40%	83,67%	81,73%	84,40%
Média 80%-20%	83,84%	83,68%	83,89%
Desvio Padrão	0,47	2,20	0,28

Na Tabela 19 são apresentados os resultados obtidos pelo modelo da árvore 3 em suas quatro proporções de treino e teste, com a aplicação da técnica de seleção de características baseada em correlação, com 8 características selecionadas (Q8/MI1/Δ*, Q8/ME2/Δ, Q8/ME3/Δ*, Q12/ME4/Δ, Q12MI5/Δ, Q12MI5/Δ*, Q16/MI5/Δ e Q16/MI5/Δ*).

Tabela 19 - Resultados obtidos pela Árvore 3 para as 2 Classes com Seleção de Características.

Árvore 3 - Com Seleção de Características									
Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade	Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade
20%-80%	Teste 1	81,00%	61,80%	88,80%	40%-60%	Teste 1	81,60%	74,10%	84,70%
	Teste 2	72,50%	78,30%	70,30%		Teste 2	76,90%	75,40%	77,50%
	Teste 3	67,80%	83,30%	62,10%		Teste 3	77,00%	74,90%	77,70%
	Teste 4	78,20%	75,30%	79,30%		Teste 4	78,30%	75,00%	79,60%
	Teste 5	71,60%	85,60%	66,40%		Teste 5	77,00%	83,60%	74,60%
	Média	74,22%	76,86%	73,38%		Média	78,16%	76,60%	78,82%
	Desvio Padrão	4,75	8,36	9,57		Desvio Padrão	1,80	3,53	3,35
60%-40%	Teste 1	78,80%	77,30%	79,40%	80%-20%	Teste 1	77,60%	68,10%	80,90%
	Teste 2	77,40%	76,10%	77,90%		Teste 2	82,20%	72,60%	86,30%
	Teste 3	80,60%	74,10%	82,80%		Teste 3	83,30%	76,30%	86,10%
	Teste 4	78,60%	69,30%	82,20%		Teste 4	83,60%	76,50%	86,50%
	Teste 5	80,80%	77,80%	81,80%		Teste 5	85,10%	77,30%	87,90%
	Média	79,24%	74,92%	80,82%		Média	82,36%	74,16%	85,54%
	Desvio Padrão	1,29	3,09	1,86		Desvio Padrão	2,55	3,44	2,40

Conforme a Tabela 19, o melhor resultado obtido foi de 82,36% de acurácia média, 74,16% de sensibilidade média e 85,54% de especificidade média na proporção 80% para treino e 20% para teste. O pior resultado encontrado foi na proporção 20% para treino e 80% para teste com acurácia média de 74,22%, sensibilidade média de 76,86% e 73,38% de especificidade média.

Analisando a Tabela 20, foi verificado que a diferença entre o melhor e o pior resultado é de 8,14% de acurácia média, possuindo um valor acima de 1 nos desvios padrões em todas as métricas de validação.

Tabela 20 - Resultados das médias das proporções obtidas pela Árvore 3 com Seleção de Características.

Árvore 3 - Com Seleção de Características			
Proporção	Acurácia	Sensibilidade	Especificidade
Média 20%-80%	74,22%	76,86%	73,38%
Média 40%-60%	78,16%	76,60%	78,82%
Média 60%-40%	79,24%	74,92%	80,82%
Média 80%-20%	82,36%	74,16%	85,54%
Desvio Padrão	2,91	1,13	4,36

Analisando as Tabela 18 e Tabela 20, verifica-se que a aplicação da técnica de seleção de características não melhorou os resultados obtidos em nenhuma das quatro proporções estabelecidas.

4.2.4 Todas as Árvores Juntas

Na Tabela 21 são apresentados os resultados obtidos pela junção dos três modelos de árvores filogenéticas, em suas quatro proporções de treino e teste, sem a aplicação da técnica de seleção de características baseada em correlação.

Tabela 21 - Resultados obtidos pela Junção de todas as Árvores para as 2 Classes.

Todas as Árvores Juntas									
Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade	Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade
20%-80%	Teste 1	80,96%	80,06%	81,31%	40%-60%	Teste 1	85,53%	83,41%	86,23%
	Teste 2	86,21%	82,85%	87,48%		Teste 2	85,41%	82,01%	86,75%
	Teste 3	86,39%	78,76%	89,24%		Teste 3	89,21%	86,46%	90,23%
	Teste 4	85,05%	73,73%	89,48%		Teste 4	89,09%	82,35%	91,74%
	Teste 5	85,94%	82,64%	87,21%		Teste 5	86,48%	82,04%	88,29%
	Média	84,91%	79,61%	86,94%		Média	87,14%	83,25%	88,65%
	Desvio Padrão	2,03	3,32	2,96		Desvio Padrão	1,68	1,68	2,08
60%-40%	Teste 1	86,65%	80,84%	89,11%	80%-20%	Teste 1	91,10%	87,50%	92,54%
	Teste 2	86,12%	84,91%	86,60%		Teste 2	88,61%	84,62%	90,15%
	Teste 3	89,68%	85,54%	91,41%		Teste 3	90,39%	92,31%	89,81%
	Teste 4	85,94%	83,54%	86,88%		Teste 4	87,90%	82,76%	90,21%
	Teste 5	88,08%	82,56%	90,51%		Teste 5	87,54%	86,25%	88,06%
	Média	87,29%	83,48%	88,90%		Média	89,11%	86,69%	90,15%
	Desvio Padrão	1,41	1,68	1,91		Desvio Padrão	1,40	3,23	1,43

Conforme a Tabela 21, o melhor resultado obtido foi de 89,11% de acurácia média, 86,69% de sensibilidade média e 90,15% de especificidade média na proporção 80% para treino e 20% para teste (melhor resultado alcançado pela metodologia proposta). O pior resultado encontrado foi na proporção 20% para treino e 80% para teste com acurácia média de 84,91%, sensibilidade média de 79,61% e 86,94% de especificidade média.

Analisando a Tabela 22, foi verificado que a diferença entre o melhor e o pior resultado é de 4,20% de acurácia média, possuindo um valor acima de 1 nos desvios padrões em todas as métricas de validação.

Tabela 22 - Resultados das médias das proporções obtidas pela Junção de todas as Árvores.

Todas as Árvores Juntas			
Proporção	Acurácia	Sensibilidade	Especificidade
Média 20%-80%	84,91%	79,61%	86,94%
Média 40%-60%	87,14%	83,25%	88,65%
Média 60%-40%	87,29%	83,48%	88,90%
Média 80%-20%	89,11%	86,69%	90,15%
Desvio Padrão	1,49	2,51	1,14

Na Tabela 23 são apresentados os resultados obtidos pela junção de todas as árvores filogenéticas, em suas quatro proporções de treino e teste, com a aplicação da técnica de seleção de características baseada em correlação, com 15 características selecionadas (Q8/MI1/Δ* e Q16/MI1/Δ referentes à Árvore 1, Q8/MI2/Δ, Q8/ME2/Δ*, Q8/ME3/Δ, Q8/MI4/Δ, Q8/MI5/Δ, Q12/MI4/Δ*, Q12/MI5/Δ, Q16/MI1/Δ*, Q16/ME3/Δ, Q16/MI4/Δ, Q16/MI4/Δ*, Q16/ME4/Δ e Q16/ME4/Δ* referentes à Árvore 3).

Tabela 23 - Resultados obtidos pela Junção de todas as Árvores para as 2 Classes com Seleção de Características.

Todas as Árvores Juntas - Com Seleção de Características									
Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade	Treino/Teste	Testes	Acurácia	Sensibilidade	Especificidade
20%-80%	Teste 1	88,60%	73,60%	94,60%	40%-60%	Teste 1	88,80%	79,40%	92,70%
	Teste 2	88,30%	79,90%	91,50%		Teste 2	88,30%	78,70%	92,20%
	Teste 3	88,60%	79,70%	91,90%		Teste 3	87,70%	81,40%	89,60%
	Teste 4	86,70%	78,20%	89,90%		Teste 4	87,00%	77,50%	90,60%
	Teste 5	87,50%	84,30%	88,80%		Teste 5	86,60%	86,70%	86,50%
	Média	87,94%	79,14%	91,34%		Média	87,68%	80,74%	90,32%
	Desvio Padrão	0,74	3,44	1,97		Desvio Padrão	0,81	3,24	2,21
60%-40%	Teste 1	89,50%	83,10%	91,90%	80%-20%	Teste 1	88,30%	79,20%	91,40%
	Teste 2	87,90%	77,40%	92,10%		Teste 2	88,30%	79,80%	91,90%
	Teste 3	87,40%	81,80%	89,30%		Teste 3	90,00%	88,80%	90,50%
	Teste 4	85,80%	71,90%	91,00%		Teste 4	88,30%	77,80%	92,50%
	Teste 5	86,50%	85,40%	86,80%		Teste 5	88,60%	80,80%	91,60%
	Média	87,42%	79,92%	90,22%		Média	88,70%	81,28%	91,58%
	Desvio Padrão	1,27	4,78	1,98		Desvio Padrão	0,66	3,88	0,66

Conforme a Tabela 23, o melhor resultado obtido foi de 88,70% de acurácia média, 81,28% de sensibilidade média e 91,58% de especificidade média na proporção 80% para treino e 20% para teste. O pior resultado encontrado foi na proporção 60% para treino e 40% para teste com acurácia média de 87,42%, sensibilidade média de 79,92% e 90,22% de especificidade média.

Analisando a Tabela 24, foi verificado que a diferença entre o melhor e o pior resultado é de apenas 1,28% de acurácia média, possuindo um valor abaixo de 1 nos desvios padrões em todas as métricas de validação.

Tabela 24 - Resultados das médias das proporções obtidas pela Junção de todas as Árvores com Seleção de Características.

Todas as Árvores Juntas - Com Seleção de Caracteristicas			
Proporção	Acurácia	Sensibilidade	Especificidade
Média 20%-80%	87,94%	79,14%	91,34%
Média 40%-60%	87,68%	80,74%	90,32%
Média 60%-40%	87,42%	79,92%	90,22%
Média 80%-20%	88,70%	81,28%	91,58%
Desvio Padrão	0,48	0,81	0,60

Analisando as Tabela 22 e Tabela 24, verifica-se que a aplicação da técnica de seleção de características obteve resultados de acurácias médias maiores em todas as proporções, exceto na proporção 80% para treino e 20% para teste.

Sintetizando os resultados obtidos entre os três modelos das árvores filogenéticas utilizados, foi observado que a árvore 2 apresentou valores maiores e mais equilibrados com sua aplicação em relação as outras duas arquiteturas de árvores devido a menor quantidade de espécies (unidade Hounsfield), pois houve a exclusão das espécies que não possuíam indivíduos (*voxel*). Outro fator observado foi que os resultados da medida de especificidade apresentaram valores mais altos do que a medida de sensibilidade, atribui-se a isso, o fato da maioria dos nódulos considerados para essa base serem da classe benigno.

A Tabela 25 apresenta os melhores e os piores resultados obtidos com a aplicação da metodologia proposta para o diagnóstico de câncer de pulmão, suas respectivas arquiteturas de árvores, além do teste realizado com a junção das árvores filogenéticas, e proporções de treino e teste. O melhor resultado de acurácia média alcançado nos testes realizados, foi com todas as árvores juntas sem a seleção de características na proporção 80% para treino e 20% para teste e o pior resultado de acurácia média foi obtido na árvore 3 com a seleção de características na proporção 20% para treino e 80% para teste.

Tabela 25 - Melhores e piores resultados obtidos com a metodologia proposta.

	Melhores Resultados				Piores Resultados			
Tipo de Árvore	Proporção	Acurácia	Sensibilidade	Especificidade	Proporção	Acurácia	Sensibilidade	Especificidade
Árvore 1	40%-60%	84,65%	80,84%	86,15%	20%-80%	82,80%	74,80%	86,03%
Árvore 1 com Seleção	80%-20%	83,92%	75,14%	87,44%	20%-80%	80,50%	72,20%	83,74%
Árvore 2	80%-20%	88,61%	85,67%	89,75%	20%-80%	85,87%	79,79%	88,26%
Árvore 2 com Seleção	20%-80%	88,28%	80,10%	91,44%	60%-40%	87,88%	81,22%	90,36%
Árvore 3	80%-20%	83,84%	83,68%	83,89%	20%-80%	82,69%	77,78%	84,67%
Árvore 3 com Seleção	80%-20%	82,36%	74,16%	85,54%	20%-80%	74,22%	76,86%	73,38%
Todas as Árvores	80%-20%	89,11%	86,69%	90,15%	20%-80%	84,91%	79,61%	86,94%
Todas as Árvores com Seleção	80%-20%	88,70%	81,28%	91,58%	60%-40%	87,42%	79,92%	90,22%

Analisando os trabalhos apresentados na literatura, observa-se que a metodologia proposta consegue resultados comparáveis aos melhores já publicados, conforme a Tabela 26.

Comparando-se os resultados para a base LIDC-IDRI, é possível observar que houve um aumento em todas as métricas de validação dos resultados na classificação de nódulos pulmonares em benigno e maligno. Outras metodologias foram comparadas, entretanto utilizam a base LIDC (84 exames) que possue uma quantidade de exames de tomografia computadorizada menor em relação à LIDC-IDRI empregada na metodologia proposta.

Levando-se em conta essas observações é possível destacar que os testes realizados no presente trabalho na tarefa de classificação de nódulos pulmonares em benigno e maligno se mostram bastante efetivos e promissores, encorajando estudos mais profundos, considerando inclusive, a utilização em conjunto com outras metodologias existentes.

Finalmente é importante salientar que, para uma comparação completamente justa das metodologias citadas seria necessário utilizar as mesmas imagens em todos os trabalhos. Outro fator que deveria ser comum às obras é a amostra utilizada, pois as metodologias deveriam usar os mesmos dados para as etapas de treinamento e teste na fase de reconhecimento de padrões.

Tabela 26 – Comparação dos resultados entre trabalhos relacionados.

		Melhores Resultados		
Trabalhos	Base Utilizada	Acurácia	Sensibilidade	Especificidade
(SILVA, 2009)	LIDC	81,00%	86,00%	76,00%
(NASCIMENTO, 2012)	LIDC	92,78%	85,64%	97,89%
(NASCIMENTO, 2012)	LIDC-IDRI	83,75%	82,95%	84,58%
Metodologia Proposta	**LIDC-IDRI**	**89,11%**	**86,69%**	**90,15%**

5. CONCLUSÃO

Os elevados índices de mortes e registros de ocorrências de câncer de pulmão no Brasil e no mundo demonstram a importância do desenvolvimento de pesquisas com o objetivo de produzir recursos para um diagnóstico precoce da doença propiciando dessa forma um tratamento mais adequado.

Com isso, o uso de ferramentas computacionais para o diagnóstico tem evoluído em técnicas e áreas de abrangência e, ainda, em interesse por parte da comunidade científica. O resultado é o uso dessas ferramentas para sugerir opiniões a médicos e especialistas e fazê-las aplicáveis cada vez mais, tornando-as presente em seu cotidiano.

Este trabalho apresentou uma metodologia CADx para o diagnóstico de nódulos pulmonares em imagens de tomografia computadorizada através da caracterização destes nódulos a partir da extração de medidas de textura obtidas com os cálculos dos Índices de Diversidade Taxonômica e Distinção Taxonômica juntamente com as abordagens de máscaras internas e externas submetidas a três arquiteturas de árvores filogenéticas. Após isso, comparou os resultados obtidos na classificação com a MVS, utilizando 4 conjuntos diferentes de distribuição de amostras para treinamento e teste, 20% e 80%, 40% e 60%, 60% e 40% e 80% e 20%.

Os resultados apresentados no Capítulo 4 demonstram o desempenho promissor da metodologia desenvolvida na tarefa de diagnóstico de nódulos pulmonares. Foram obtidos valores de 86,69% de sensibilidade média, 90,15% de especificidade média e 89,11% de acurácia média no teste que juntou todas as árvores filogenéticas.

É importante destacar que uma das dificuldades do trabalho foi em relação à própria base de dados utilizada, pelo fato de possuírem diagnósticos realizados de forma subjetiva, com anotação feita por diferentes radiologistas. No entanto, ainda assim, existem fatores que indicam a necessidade da ampliação e desenvolvimento deste trabalho. Nesse sentido são colocados alguns aspectos de melhoria:

- Avaliação de outros índices de diversidade na tarefa de classificação dos nódulos.

- Investigação de outras técnicas de seleção de características, como Análise de Componentes Principais (PCA), visando alcançar melhores resultados.

- O classificador MVS utilizado neste trabalho pode ser substituído por outros classificadores com o objetivo de avaliar seu desempenho na tarefa de reconhecimento de padrões maligno e benigno em regiões extraídas dos nódulos pulmonares.

- Utilização de outras bases de imagens de TC para melhorar a validação da metodologia.

Por fim, a metodologia apresentada neste trabalho poderá integrar uma ferramenta CAD/CADx a ser aplicada na detecção e diagnóstico de câncer pulmonar, no intuito de classificar os nódulos em maligno e benigno. Dessa forma tornando mais ágil e menos exaustiva a análise de exames pelos especialistas.

REFERÊNCIAS

ABC.MED.BR. **Tomografia computadorizada. Como é o exame?**. Disponível em: <http://www.abc.med.br/p/exames-e-procedimentos/344744/tomografia-computadorizada-como-e-o-exame.htm>. Acesso em: 4 dez. 2014.

ARAÚJO, G. S. de. **Filogenia de Proteomas**. 2003. Tese de Doutorado. Universidade Federal de Mato Grosso do Sul.

ARMATO III, S. G. et al. **The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans.** Medical physics, v. 38, n. 2, p. 915-931, 2011.

BLAND, M. et al. **An introduction to medical statistics**. Oxford University Press, 2000.

BRAGA, J. L. **Previsão de Mortalidade, Tempo de Estadia e Tempo de Ventilação Mecânica em UTI de Hospital Particular do Grande Recife Utilizando Técnicas de Aprendizado de Máquina.** Monografia apresentada ao curso de Engenharia da Computação da Universidade de Pernambuco. Recife, p. 73. 2005.

CARVALHO, A. O. et al **Classificação de Nódulo pulmonar baseado em tomografia computadorizada usando Índice de Diversidade Taxonômico e SVM**. São Luís: Preprint Submitted to Pattern Recognition, February 17, 2014.

CLARKE, K. R.; WARWICK, R. M. **The taxonomic distinctness measure of biodiversity: weighting of step lengths between hierarchical levels.** Marine Ecology Progress Series, v. 184, p. 21-29, 1998.

CONCI, A.; AZEVEDO, E.; LETA, F. R. **Computação Gráfica**. Vol. 2-Processamento de Imagens Digitais. 2008.

CHANG, C.C; LIN, C.J. **LIBSVM – A Library for Support Vector Machines**, 2003. Disponível em: <http://www.csie.ntu.edu.tw/cjlin/libsvm/>.

CHATE, R. C.; FUNARI, M. B. de G.. **Nódulo pulmonar; Lung nodule**. RBM rev. bras. med, v. 68, n. 1/2, 2011.

CHAVES, A. da C. F. **Extração de regras fuzzy para Máquinas de vetor de Suporte (SVM) para classificação em múltiplas classes.** Rio de Janeiro: Tese apresentada como

requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio, 2006.

FACON, J. **Processamento e Análise de Imagens**. Curso e Mestrado em Informática Aplicada. Pontifícia Universidade Católica do Paraná. p. 128. 2002.

FREIRE, T. P. **Classificação em nódulos e não nódulos baseado em imagens de tomografia computadorizada usando índices de diversidade e máquina de vetor de suporte**. Dissertação (Mestrado em Engenharia de Eletricidade). Universidade Federal do Maranhão 2014.

GIBBONS, J. D.; KOTZ, S.; JOHNSON, N. L. **Encyclopedia of Statistical Science**. Encyclopedia of Statistical Science, v. 7, 1988.

GONZALEZ, R. C.; WOODS, R. E. **Processamento de imagens digitais**. Edgard Blucher, 2002.

GORENSTEIN, M. R. **Diversidade de espécies em comunidades arbóreas: aplicação de índices de distinção taxonômica em três formações florestais do estado de São Paulo**. Dissertação (Recursos Naturais) - Universidade de São Paulo, Piracicaba, SP, 2009.

HALL, M. et al. The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, v. 11, n. 1, p. 10-18, 2009.

HALL, M. A.; SMITH, L. A. **Practical feature subset selection for machine learning.** 1998.

HARALICK, Robert M.; SHANMUGAM, Karthikeyan; DINSTEIN, Its' Hak. **Textural features for image classification.** Systems, Man and Cybernetics, IEEE Transactions on, n. 6, p. 610-621, 1973.

HAYKIN, S. S. **Redes neurais**. Bookman, 2001.

INCA - INSTITUTO NACIONAL DE CÂNCER. **O que causa o câncer?,** 2014. Disponível em: <http://www.inca.gov.br/conteudo_view.asp?id=81>. Acesso em: 03 dez. 2014.

INCA - INSTITUTO NACIONAL DE CÂNCER. **Tipos de Câncer: Pulmão**, 2014. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/pulmao>. Acesso em: 03 dez. 2014.

JABON, S. A.; RAICU, D. S.; FURST, J. D. **Content-based versus semantic-based retrieval: an LIDC case study.** In: SPIE Medical Imaging. International Society for Optics and Photonics, 2009. p. 72631L-72631L-8.

KOLLER, D.; SAHAMI, M.. **Toward optimal feature selection.** 1996.

KUMAR, S. A. et al. **Robust and automated lung nodule diagnosis from ct images based on fuzzy systems.** In: Process Automation, Control and Computing (PACC), 2011 International Conference On. IEEE, 2011. p. 1-6.

MAGURRAN, A. E. **Measuring biological diversity**. 2004.

NASCIMENTO, L. B. **Classificação de Nódulos Pulmonares em Maligno e Benigno utilizando os Índices de Diversidade de Shannon e de Simpson**. Dissertação (Mestrado em Engenharia de Eletricidade). Universidade Federal do Maranhão 2012.

NOGUEIRA, A. et al. **Um Overview Sobre Reconhecimento de Padrões.** II Simpósio de Excelência em Gestão e Tecnologia, Rio de Janeiro, 2007.

OLIVEIRA, F. S. S. de. **Classificação de Tecidos da Mama em Massa e Não-Massa usando Índice de Diversidade Taxonômico e Máquina de Vetores de Suporte.** Dissertação (Mestrado em Engenharia de Eletricidade). Universidade Federal do Maranhão 2013.

PAPPA, G. L. ***Seleção de atributos utilizando Algoritmos Genéticos multiobjetivos*.** Diss. Pontifícia Universidade Católica do Paraná, 2002.

PARVEEN, S. S; C. KAVITHA. **Classification of Lung Cancer Nodules using SVM Kernels.** International Journal of Computer Applications 95 (2014).

PEDRINI, H.; SCHWARTZ, W. R. **Análise de imagens digitais: princípios, algoritmos e aplicações**. Thomson Learning, 2008.

PIANKA, E. R. **Evolutionary ecology**. Eric R. Pianka, 1994.

RICOTTA, C.. **A parametric diversity measure combining the relative abundances and taxonomic distinctiveness of species.** Diversity and Distributions, v. 10, n. 2, p. 143-146, 2004.

SANTOS, V. K. **Uma generalização da distribuição do índice de diversidade generalizado por good com aplicação em ciências agrárias**. 2009. 56 f. Dissertação (Biometria e Estatística Aplicada) - Universidade Federal de Pernambuco, Recife, PE, 2009.

SILVA, C. A. da. **Caracterização de nódulos pulmonares solitários utilizando índice de Simpson e máquina de vetores de suporte.** 2009. Dissertação (Engenharia Elétrica) - Universidade Federal do Maranhão, São Luis, MA, 2009.

SILVA, I. A. da; BATALHA, M. A. **Taxonomic distinctness and diversity of a hyperseasonal savanna in central Brazil**. Diversity and distributions, v. 12, n. 6, p. 725-730, 2006.

SOUSA, U. S. **Classificação de massas na mama a partir de imagens mamográficas usando o índice de diversidade de shannon-wiener**. 2011. 69 f. Dissertação (Engenharia Elétrica) - Universidade Federal do Maranhão, São Luis, MA, 2011.

TARTAR, A., A. AKAN, and N. KILIC. **A novel approach to malignant-benign classification of pulmonary nodules by using ensemble learning classifiers.** *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014.

VANDAMME, P. et al. **Polyphasic taxonomy, a consensus approach to bacterial systematics**. Microbiological reviews, v. 60, n. 2, p. 407-438, 1996.

VAPNIK, V. N.; VAPNIK, V. **Statistical learning theory**. New York: Wiley, 1998.

VIANA, G. V. R. CEARÁ, Fortaleza. **Técnicas para construção de árvores filogenéticas.** Fortaleza: UFCE, 2007.