

UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE

LEONARDO BARROS NASCIMENTO

**CLASSIFICAÇÃO DE NÓDULOS PULMONARES EM MALIGNO
E BENIGNO UTILIZANDO OS ÍNDICES DE DIVERSIDADE
DE SHANNON E DE SIMPSON**

São Luís

2012

LEONARDO BARROS NASCIMENTO

**CLASSIFICAÇÃO DE NÓDULOS PULMONARES EM MALIGNO
E BENIGNO UTILIZANDO OS ÍNDICES DE DIVERSIDADE
DE SHANNON E DE SIMPSON**

Dissertação de Mestrado submetida à Coordenação do Programa de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão como parte dos requisitos para obtenção do título de Mestre em Engenharia de Eletricidade na área de Ciência da Computação.

Orientador: Prof. Anselmo Cardoso de Paiva
Co-orientador: Prof. Aristófanés Corrêa Silva

**São Luís
2012**

Nascimento, Leonardo Barros.

Classificação de nódulos pulmonares em maligno e benigno utilizando os Índices de Diversidade de Shannon e Simpson / Leonardo Barros Nascimento – São Luís, 2012.

59 f.

Orientador: Anselmo Cardoso de Paiva.

Co-orientador: Aristófanês Corrêa Silva.

Dissertação (Mestrado) – Universidade Federal do Maranhão, Programa de Pós-Graduação em Engenharia de Eletricidade, 2012.

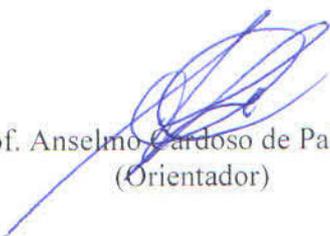
1. Diagnóstico de nódulo pulmonar – Índice de Diversidade de Shannon e Simpson. 2. MVS. 3. CADx. I. Título.

CDU 621.386.84:612.24-073

**CLASSIFICAÇÃO DE NÓDULOS PULMONARES EM
MALIGNO E BENIGNO UTILIZANDO OS ÍNDICES DE
DIVERSIDADE DE SHANNON E DE SIMPSON**

Leonardo Barros Nascimento

Dissertação aprovada em 20 de abril de 2012.



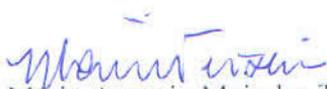
Prof. Anselmo Cardoso de Paiva, Dr.
(Orientador)



Prof. Aristófares Corrêa Silva, Dr.
(Co-Orientador)



Prof. Cristina Nader Vasconcelos, Dra.
(Membro da Banca Examinadora)



Prof. Mário Antonio Meireles Teixeira, Dr.
(Membro da Banca Examinadora)

À minha família e amigos.

AGRADECIMENTOS

À Deus pela saúde e pelo dom da vida.

À meus pais Raimunda e Valdeir e minha irmã Auricleide pela ajuda e incentivo.

À minha namorada Ana Lúcia pela compreensão e apoio durante a caminhada de realização deste trabalho.

Ao meu orientador Anselmo, pela paciência, competência e dedicação.

Ao meu co-orientador Ari, pela disponibilidade, conselhos e sugestões.

Aos demais professores, Zair e Denivaldo, pela contribuição no crescimento acadêmico.

Aos meus colegas, Fernando, Péterson e Janilson, pela força e por compartilharem conhecimento.

À Fapema pelo apoio financeiro.

À todos que de alguma forma contribuíram para a realização deste trabalho.

*“A forma mais terrível de naufrágio é não
partir.”*

Amyr Klink

RESUMO

O câncer de pulmão é ainda a maior causa de mortalidade por câncer em todo mundo, com uma das menores taxas de sobrevivência a partir do diagnóstico. Por isso, sua detecção precoce é importante para aumentar as chances de cura do paciente, e de quanto mais informações o médico dispuser, mais preciso será o diagnóstico. Diante do exposto, o presente trabalho apresenta uma metodologia de caracterização de nódulos pulmonares, objetivando se tornar uma ferramenta computacional utilizada para sugerir sobre a malignidade ou benignidade dos mesmos, atuando como uma segunda opinião junto ao especialista. A metodologia foi aplicada em duas bases de dados diferentes, uma com 73 nódulos, sendo 26 malignos e 47 benignos, e outra com 1034 nódulos sendo 517 malignos e 517 benignos. Os Índices de Diversidade de Shannon e de Simpson foram utilizados como descritores de textura. As características geradas foram submetidas à etapa de seleção de características com a utilização da Análise Discriminante *stepwise*. Após essa etapa foi realizada a classificação pela Máquina de Vetores de Suporte (MVS) onde foram obtidas taxas de sensibilidade de 85,64%, especificidade de 97,89% e acurácia de 92,78%.

Palavras-chave: Diagnóstico de Nódulo Pulmonar, Índices de Diversidade de Shannon e de Simpson, SVM, CADx.

ABSTRACT

Lung cancer is still 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 curing the patient. The diagnosis is more accurate if the specialist has more information. In view of the above, this work presents a methodology for characterization about the malignancy or benignity of pulmonary nodules, acting as a second opinion for the expert. The methodology was applied in two different databases, one with 73 nodes, 26 malignant and 47 benign, and other with 1034 nodes and 517 malignant and 517 benign. The Diversity Indices of Shannon and Simpson were used as texture descriptors. The features generated were then subjected to the step of feature selection using the stepwise Discriminant Analysis. After this stage, they were classified by the Support Vector Machine (SVM) where we obtained sensitivity of 85.64%, specificity of 97.89% and accuracy of 92.78%.

Keywords: Diagnosis of Pulmonary Nodule, Shannon and Simpson Diversity Indices, SVM, CADx.

Artigos científicos aceitos para publicação

NASCIMENTO, L. B., PAIVA, A. C., SILVA, A. C. Lung Nodules Classification in CT Images Using Shannon and Simpson Diversity Indices and SVM. International Conference on Machine Learning and Data Mining (MLDM), Berlin, 2012.

SUMÁRIO

1	INTRODUÇÃO	14
1.1	Objetivos	15
1.2	Trabalhos relacionados	15
1.3	Estrutura do trabalho	17
2	FUNDAMENTAÇÃO TEÓRICA	19
2.1	Nódulo Pulmonar Solitário	19
2.2	Sistema de Auxílio à Detecção e Diagnóstico	20
2.3	Processamento de Imagens Digitais	21
2.3.1	Equalização de histograma	23
2.3.2	Quantização	25
2.4	Análise de Textura	26
2.4.1	Índices de Diversidade	27
2.4.1.1	Índice de Shannon	27
2.4.1.2	Índice de Simpson	28
2.5	Seleção de Características	29
2.5.1	Análise Discriminante Linear <i>Stepwise</i>	29
2.6	Reconhecimento de Padrões	30
2.6.1	Máquina de Vetores de Suporte	31
2.7	Avaliação dos Resultados	34
3	METODOLOGIA	36
3.1	Aquisição das Imagens	36
3.1.1	Base I: LIDC	37
3.1.2	Base II: LIDC-IDRI	39
3.2	Pré-Processamento	40
3.3	Segmentação 3D dos Nódulos	40
3.4	Extração de Características	40
3.4.1	Características de Textura	40
3.5	Seleção de características	42
3.6	Classificação	42
3.7	<i>Software e Hardware</i> Utilizados	42
4	TESTES E RESULTADOS	43
4.1	Base I: LIDC	44
4.1.1	Testes com o Índice de Shannon	44
4.1.1.1	Abordagem em anéis	44
4.1.1.2	Abordagem em esferas	45
4.1.2	Testes com o Índice de Simpson	45
4.1.2.1	Testes utilizando abordagem em anéis	45

4.1.2.2.	Testes utilizando abordagem em esferas	46
4.1.3	Testes com o Índice de Shannon e de Simpson	47
4.2	Base II: LIDC-IDRI	47
4.2.1	Testes com o Índice de Shannon	47
4.2.1.1.	Abordagem em anéis	48
4.2.1.2.	Abordagem em esferas.....	48
4.2.2	Testes com o Índice de Simpson.....	48
4.2.2.1.	Testes utilizando abordagem em anéis	49
4.2.2.2.	Testes utilizando abordagem em esferas	49
4.2.3	Testes com o Índice de Shannon e Simpson	50
4.2.3.1.	Testes utilizando abordagem em anéis	50
4.2.3.2.	Testes utilizando abordagem em esferas	50
4.3	Avaliação Final	51
5	CONCLUSÃO	54
	REFERÊNCIAS	55

LISTA DE FIGURAS

Figura 2.1 - Exemplo de nódulo pulmonar. Fonte: (ARMATO III, MCLENNAN, <i>et al.</i> , 2011)	19
Figura 2.2 - Convenção dos eixos para representação de imagens digitais. Fonte: (GONZALEZ e WOODS, 2000).....	21
Figura 2.3 - Passos fundamentais para processamento de imagens digitais. Fonte: (GONZALEZ e WOODS, 2000).....	22
Figura 2.4 - Equalização de Histograma. (a) Imagem Original, (b) Imagem Equalizada, (c) Histograma Original e (d) Histograma Equalizado. Adaptado de (MATHWORKS, 2012).....	24
Figura 2.5 – Níveis de quantização e gráfico da função de quantização. Fonte: (GOMES e VELHO, 1994)	25
Figura 2.6 - Separação de duas classes através de hiperplanos.....	32
Figura 2.7 - Vetores de suporte (destacados por círculos).	33
Figura 3.1- Estágios da Metodologia.....	36
Figura 3.2 - Ilustração do resumo dos nódulos. Adaptado de (HORSTHEMKE, RAICU e FURST, 2008).	38
Figura 3.3 - Ilustração do resumo do diagnóstico dos nódulos. Adaptado de (JABON, RAICU e FURST, 2009).....	39
Figura 3.4 - Representação em 2D da análise em esferas concêntricas aplicada ao nódulo. ...	41
Figura 3.5 - Representação em 2D da análise em anéis esféricos aplicada ao nódulo.....	41

LISTA DE TABELAS

Tabela 1 - Resultados dos testes utilizando o Índice de Shannon, em anéis, extraído da Base I.	44
Tabela 2 - Resultados dos testes utilizando o Índice de Shannon, em esferas, extraído da Base I.....	45
Tabela 3 - Resultados dos testes utilizando o Índice de Simpson, em anéis, extraído da Base I.	46
Tabela 4 - Resultados dos testes utilizando o Índice de Simpson, em esferas, extraído da Base I.....	46
Tabela 5 - Resultados dos testes utilizando os Índices de Shannon e Simpson, em esferas, extraídos da Base I.....	47
Tabela 6 - Resultados dos testes utilizando o Índice de Shannon, em anéis, extraído da Base II.	48
Tabela 7 - Resultados dos testes utilizando o Índice de Shannon, em esferas, extraído da Base II	48
Tabela 8 - Resultados dos testes utilizando o Índice de Simpson, em anéis, extraído da Base II.	49
Tabela 9 - Resultados dos testes utilizando o Índice de Simpson, em esferas, extraído da Base II	49
Tabela 10 - Resultados dos testes utilizando os Índices de Shannon e Simpson, em anéis, extraídos da Base II	50
Tabela 11 - Resultados dos testes utilizando os Índices de Shannon e Simpson, em esferas, extraídos da Base II.	50
Tabela 12 - Resultados dos testes utilizando os Índices de Shannon e Simpson, em anéis e esferas, extraídos da Base II.	51
Tabela 13 - Resumo dos melhores resultados obtidos pela metodologia em cada índice de diversidade isoladamente e em conjunto, para cada base de dados.....	52
Tabela 14 – Comparação dos resultados entre trabalhos relacionados.	53

LISTA DE ABREVIATURAS E SIGLAS

ADL	- Análise Discriminante Linear
DICOM	- Digital Imaging Communications in Medicine
CAD	- Computer-aided Detection
CADx	- Computer-aided Diagnosis
FN	- Falso Negativo
FP	- Falso Positivo
IDRI	- Image Database Resource Initiative
INCA	- Instituto Nacional do Câncer
LIDC	- Lung Image Database Consortium
MVS	- Máquina de Vetores de Suporte
NCI	- National Cancer Institute
NEMA	- National Electrical Manufacturers Association
VP	- Verdadeiro Positivo
VN	- Verdadeiro Negativo
XML	- eXtensible Markup Language

1 INTRODUÇÃO

Câncer é o nome dado a um conjunto de mais de 100 doenças que têm em comum o crescimento desordenado de células que invadem tecidos e órgãos. Dividindo-se rapidamente, estas células tendem a ser muito agressivas e incontroláveis, determinando a formação de tumores malignos, que podem espalhar-se para outras regiões do corpo. 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, 2011).

O surgimento do câncer depende da intensidade e duração da exposição das células aos agentes causadores. Por exemplo, o risco de uma pessoa desenvolver câncer de pulmão é diretamente proporcional ao número de cigarros fumados por dia e ao número de anos que ela se submete ao vício (INCA, 2011).

O câncer de pulmão é o mais comum de todos os tumores malignos, apresentando um aumento por ano de 2% na sua incidência mundial. Em 90% dos casos diagnosticados está associado ao consumo de derivados de tabaco (INCA, 2011).

No Brasil as estimativas de novos casos de câncer de pulmão para 2012 são de 27.320, sendo 17.210 homens e 10.110, mulheres (INCA, 2012).

A maneira mais fácil de diagnosticar o câncer de pulmão é por meio de Raio-X do tórax, complementado por uma Tomografia Computadorizada (TC). Essas tecnologias auxiliam os especialistas. Por meio das imagens obtidas, é possível desenvolver sistemas assistidos por computador que auxiliem na detecção do nódulo, e que podem até fornecer uma possível indicação de diagnóstico, funcionando assim como 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 assistido por computador).

Os CAD são sistemas que 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). Esses sistemas utilizam técnicas de processamento de imagens para auxiliar na realização do diagnóstico.

1.1 Objetivos

Este trabalho tem como principal objetivo propor uma metodologia para diagnóstico de nódulos pulmonares, por meio de análise de imagens de tomografia computadorizada, de forma a determinar padrões de comportamento de benignidade e malignidade dos nódulos, utilizando medidas de textura obtidas com o cálculo de índices de diversidade: Índice Diversidade de Shannon (SHANNON e WEAVER, 1949) e de Simpson (SIMPSON, 1949). 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 nos índices de diagnósticos corretos.

1.2 Trabalhos relacionados

Muitas metodologias computacionais têm sido desenvolvidas para a tarefa de detecção e diagnóstico a partir de imagens de Tomografia Computadorizada. Em (SANTOS, 2011) é apresentado um método de detecção de nódulos pulmonares pequenos utilizando Modelo de mistura Gaussiana e Matriz Hessiana sobre a base pública LIDC (ARMATO III, MCLENNAN, *et al.*, 2004), o trabalho alcança uma sensibilidade de 80,5%, no entanto é apurada uma taxa relativamente alta de 1,17 falsos positivos por fatia e há casos em que a detecção falha como em alguns nódulos conectados à borda da caixa torácica.

Já em (LIU, YANG e ZHAO, 2009), foi proposto um esquema de detecção de nódulo pulmonar auxiliado por computador, baseado na análise de *voxel* (autovalores da Matriz Hessiana) em imagens de Tomografia Computadorizada. O método possui várias etapas, dentre elas, a segmentação de regiões do pulmão e melhoramento dos candidatos iniciais a nódulo. Em seu trabalho, o autor utilizou a Máquina de Vetores de Suporte (MVS) em 32 casos das bases LIDC e NSRCT-LUNG para a classificação de *voxels* de nódulo e não nódulo, sendo que nesse estágio é empregada ainda regra de decisão sobre a predição da MVS. Na avaliação de performance da MVS e regra de decisão, os resultados obtidos chegaram a valores de sensibilidade de 0.9375, acurácia de 0.8782 e especificidade de 0.8766,

no entanto, são gerados muitos falsos positivos na fase inicial, de forma que o próprio autor sugere a utilização de alguma técnica adicional a ser introduzida na metodologia.

Em (CAMPOS, 2009) é proposto um novo algoritmo de segmentação baseado em crescimento de regiões para detecção de nódulos pulmonares em imagens de tomografia computadorizada. Resultados do experimento utilizando 61 exames fornecidos pelas bases LIDC e HUPE, chegaram a uma sensibilidade de 80,9 % com ocorrência de cerca de 0,23 falsos positivos por fatia, entretanto a técnica de pré-segmentação utilizada no trabalho exclui nódulos pequenos do restante do processo.

Em (EL-BAZ, GIMEL'FARB, *et al.*, 2010) e (EL-BAZ, NITZKEN, *et al.*, 2011) são propostos métodos de diagnóstico de nódulos pulmonares malignos, utilizando uma base proprietária contendo 109 nódulos (51 malignos e 58 benignos). O emprego de características morfológicas em (EL-BAZ, GIMEL'FARB, *et al.*, 2010) resultou em 94,4% de classificação correta e a utilização da distribuição espacial de intensidades das imagens em (EL-BAZ, NITZKEN, *et al.*, 2011) resultou em 96,3% de precisão na classificação, mostrando que as metodologias apresentam potencial para serem mais bem investigadas.

Em (LEE, KOUZAN e NASIERDING, 2009) é proposto um sistema automatizado de detecção de nódulo pulmonar. O sistema emprega florestas aleatórias como classificador base, e utiliza uma arquitetura única para classificação auxiliada por agrupamento. Utilizando 5721 imagens de TC da base LIDC, os resultados experimentais alcançados pelo sistema obtiveram taxas de sensibilidade de 97,92%, especificidade de 96,28% e 97,10% de acurácia, no entanto a etapa de classificação da metodologia, por vezes, possui pior desempenho, em tempo de execução, em relação a outros classificadores para algumas configurações de treino / teste.

Em (KUMAR, RAMESH, *et al.*, 2011) é descrito um sistema que pode detectar e diagnosticar nódulos pulmonares em malignos ou benignos 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. O objetivo do trabalho é também diminuir a quantidade de falsos positivos mantendo as boas taxas de verdadeiros positivos. Utilizando uma base proprietária com 40 casos, o método alcança bons resultados com acurácia de 90% com 0,3 falsos positivos por fatia. No entanto, a metodologia deve ser testada mais amplamente, uma vez que os trabalhos comparativos dos resultados,

embora inferiores apresentam, por vezes, testes com um número bem maior de casos ou mesmo utilizam imagens de radiografia convencional.

Índices de diversidade como descritores de textura têm sido usados para fornecer informação discriminante entre nódulos pulmonares. Em (SILVA, 2009) o índice de Simpson em conjunto com as características de desproporção esférica, densidade esférica e esfericidade presentes em (SOUZA, 2007) foi proposto para classificação quanto à natureza dos nódulos (maligna e benigna) utilizando as bases LIDC e HUPE e o classificador MVS, alcançando resultados de 90% de sensibilidade, 96,67% de especificidade e 95% de acurácia.

Vários trabalhos relacionados apresentam boas taxas tanto para a detecção quanto para o diagnóstico de nódulos pulmonares em imagens de TC. No entanto, ainda é necessário identificar técnicas que permitam melhorar e consolidar estes resultados. Podemos verificar 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.

Um destaque importante nos trabalhos é o uso de bases de imagens públicas, como LIDC e HUPE, onde a LIDC ganha destaque em quantidade de trabalhos onde é utilizada. Essa base possui um bom número de exames, além de informações extras associadas. Devido às boas referências este banco também foi utilizado neste trabalho.

É também possível destacar a utilização do classificador Máquina de Vetores de Suporte nos trabalhos relacionados. Esse método é bastante utilizado em aplicações de reconhecimento de padrões e na sub-área de imagens médicas, apresentando bons resultados. Ressalta-se ainda a utilização dos exames em 3D com exceção do trabalho de (KUMAR, RAMESH, *et al.*, 2011).

1.3 Estrutura do trabalho

Nos próximos capítulos a metodologia proposta será apresentada com maiores detalhes. Esse conteúdo está organizando em mais quatro capítulos, cuja descrição sucinta é apresentada a seguir.

No Capítulo 2, apresentam-se brevemente algumas propriedades dos nódulos pulmonares. Apresenta-se ainda a fundamentação teórica relativa aos conteúdos em

processamento digital de imagens, que foram necessários no trabalho. Aborda-se, em seguida, a extração de textura através dos Índices de Diversidade de Shannon e de Simpson, seleção de variáveis com Análise Discriminante *stepwise* e a técnica de classificação denominada Máquinas de Vetores de Suporte.

No Capítulo 3 descreve-se o método para realizar a classificação de nódulos pulmonares em benigno e maligno, extraídos de exames de tomografias computadorizadas, utilizando a extração de textura baseada nos índices de Shannon e de Simpson e Máquina de Vetores de Suporte.

No Capítulo 4 os resultados obtidos através da aplicação da metodologia proposta são apresentados e discutidos.

Por fim, no Capítulo 5, apresentam-se as conclusões sobre o trabalho e sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica utilizada no desenvolvimento deste trabalho e necessária para compreensão dos métodos utilizados para alcançar os objetivos desejados. Aborda-se o Nódulo Pulmonar Solitário, teoria sobre processamento de imagens, métodos de análise de textura baseado em Índice de Diversidade com o uso do Índice de Shannon e de Simpson, seleção de características utilizando Análise Discriminante *stepwise*, classificação e reconhecimento de padrão utilizando Máquina de Vetores de Suporte e técnicas de validação de resultados.

2.1 Nódulo Pulmonar Solitário

O Nódulo Pulmonar Solitário (NPS) 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. A Figura 2.1 mostra um exemplo de nódulo pulmonar em uma fatia de exame de TC.

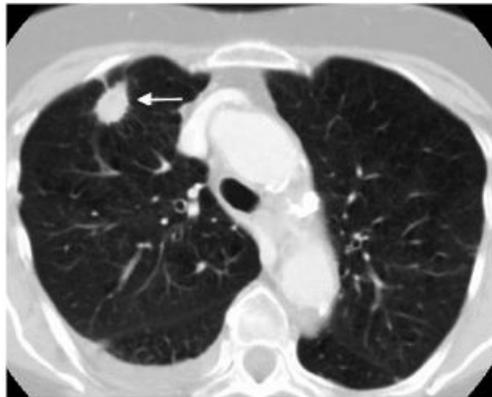


Figura 2.1 - Exemplo de nódulo pulmonar. Fonte: (ARMATO III, MCLENNAN, *et al.*, 2011)

Dentre as causas do aparecimento dos NPS temos as causas neoplásicas e as não neoplásicas. Podemos citar como exemplos de causas não-neoplásicas: Granulomas¹ infecciosos (tuberculose, histoplasmose) ou não infecciosos (sarcoidose, granulomatose de Wegener); Hamartomas²; doenças parasitárias; mal formações arteriovenosas. Entre as causas

¹ Termo aplicado a lesões inflamatórias, nodulares, em geral pequenas, granulares, firmes e persistentes.

² Tumor resultante de novo crescimento de tecidos normais.

neoplásicas podemos citar: tumores broncogênicos (carcinomas e linfoma primário de pulmão); tumores metastáticos como o sarcoma de Kaposi e os adenocarcinomas de qualquer origem.

No contexto do diagnóstico por imagem dos NPS a tomografia computadorizada possui várias vantagens sobre a radiografia simples, dentre elas podemos citar: melhor resolução, capacidade de detecção de nódulos pequenos (com 3-4 mm), melhor caracterização dos aspectos morfológicos do NPS (p.ex.: a determinação de gordura dentro do nódulo é um indicador de benignidade), etc.

Ainda nesse contexto, algumas das características dos nódulos pulmonares que ajudam a inferir sobre a probabilidade de benignidade e malignidade incluem (CHATE e 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 freqüentes 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.

Contudo, o diagnóstico definitivo de malignidade é dado apenas pelo exame citopatológico³ do material obtido por procedimentos que estão se tornando de menor morbidade, como a biópsia transbrônquica e transtorácica.

2.2 Sistema de Auxílio à Detecção e Diagnóstico

A avaliação do radiologista é baseada em subjetividade, estando sujeita a variações intra

³Exame feito para estudar as alterações celulares, principalmente as do núcleo das células.

e interpessoais, bem como perda de informação devido à natureza sutil do achado radiológico, baixa qualidade da imagem, sobreposição de estruturas, fadiga visual ou distração. Desta forma, os sistemas de Auxílio à Detecção de Diagnóstico podem se tornar ferramentas úteis no auxílio a essa avaliação. Já foi demonstrado, por exemplo, que uma dupla leitura (por dois radiologistas) pode aumentar a sensibilidade do diagnóstico (AZEVEDO-MARQUES, 2001). Deste modo, os sistemas de auxílio podem atuar justamente como um segundo especialista, ou uma segunda opinião para a tomada de decisão, com a finalidade de melhorar a consistência da interpretação da imagem radiológica.

Basicamente existem dois tipos de aplicações desses sistemas: Um é o auxílio à detecção de lesões (CAD) e o outro é o auxílio ao diagnóstico de lesões (CADx). Pode-se desenvolver essas aplicações para todos os tipos de exame de todas as partes do corpo, como crânio, tórax, abdome, osso e sistema vascular, entre outros. Porém, os principais objetos de pesquisa para o desenvolvimento de sistemas CAD e CADx têm sido as áreas de mamografia, tomografia de tórax, e angiografia.

2.3 Processamento de Imagens Digitais

Uma imagem pode ser descrita por uma função $f(x,y)$ da intensidade luminosa, sendo seu valor, em qualquer ponto de coordenadas espaciais (x,y) , proporcional ao brilho da imagem naquele ponto (Figura 2.2).

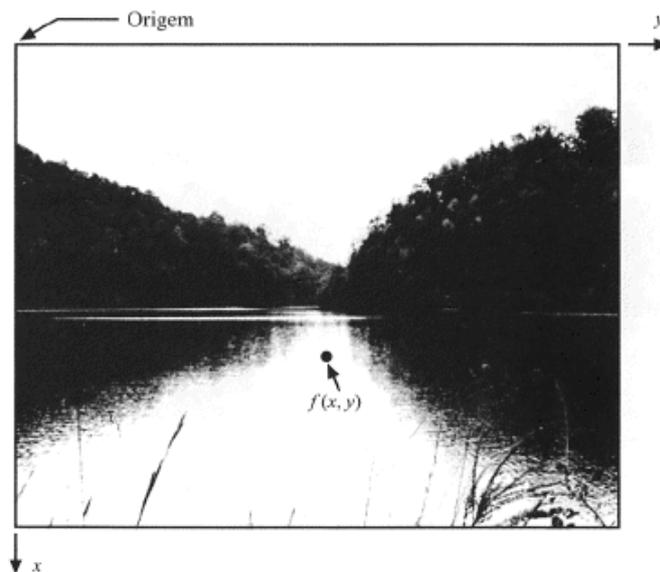


Figura 2.2 - Convenção dos eixos para representação de imagens digitais. Fonte: (GONZALEZ e WOODS, 2000).

Já no caso de uma imagem que possui informações em intervalos ou bandas distintas de frequência, é necessário uma função $f(x,y)$ para cada banda. É o caso de imagens coloridas padrão RGB, que são formadas pela informação de cores primárias, como o vermelho, verde e azul. Para o processamento da imagem digitalizada, é fundamental representar sua informação num formato adequado ao tratamento computacional.

Uma imagem pode ser representada por uma matriz, em que os índices de linha e coluna referenciam o brilho médio amostrado no ponto correspondente da cena, e cada elemento chama-se pixel (abreviação de "picture element"), sendo codificado com R níveis de cinza. Em imagens 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 abrange uma ampla escala de *hardware*, *software* e fundamentos teóricos e esse processo segue várias etapas. O fluxograma padrão dessas etapas está representado na Figura 2.3.

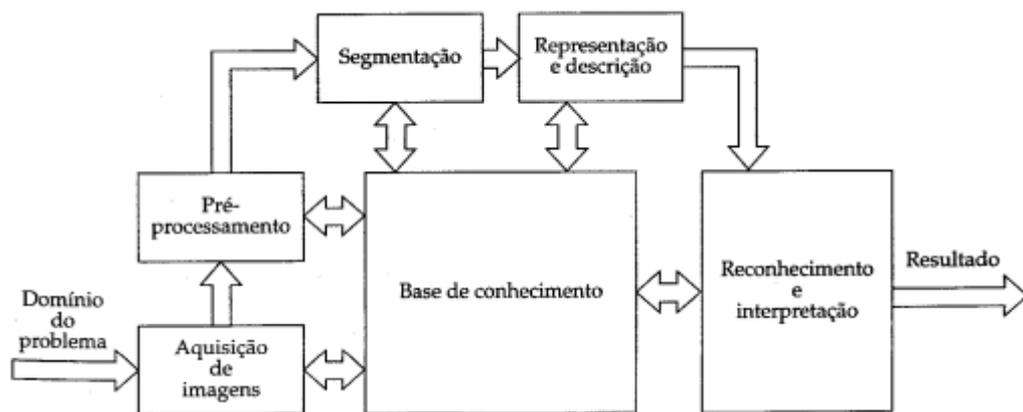


Figura 2.3 - Passos fundamentais para processamento de imagens digitais. Fonte: (GONZALEZ e WOODS, 2000).

A Figura 2.3 mostra que o objetivo é produzir um resultado a partir do domínio do problema por meio de processamento de imagens. No presente trabalho, o domínio do problema consiste em nódulos pulmonares e o objetivo é classificá-los em maligno e benigno. Assim a saída será correspondente a rótulos referente ao diagnóstico dos nódulos.

O primeiro passo então é a aquisição da imagem, isto é, adquirir uma imagem digital produzida por um sensor de imageamento, como um tomógrafo por exemplo. Concluída essa etapa, o próximo passo trata de pré-processar a imagem. A função chave no pré processamento é melhorar a imagem de forma a aumentar as chances de sucesso dos

processos seguintes. Neste trabalho o pré-processamento consiste na equalização do histograma das imagens de TC.

O próximo estágio trata da segmentação. Em termos gerais, a segmentação divide uma imagem de entrada em partes ou objetos constituintes. Um procedimento de segmentação robusto favorece substancialmente a solução bem sucedida de um problema de imageamento. Para este trabalho a segmentação é definida por especialistas que delimitam manualmente os nódulos pulmonares.

Na representação e descrição pretende-se transformar os dados iniciais numa forma adequada para o subsequente processamento computacional. Nessa fase são extraídas características que resultem em alguma informação quantitativa de interesse para discriminação entre classes. Em se tratando de diagnóstico de nódulos pulmonares, a descrição das características texturais é um dado importante para essa tarefa.

O último estágio envolve reconhecimento e interpretação. Reconhecimento é o processo que atribui um rótulo a um objeto, baseado na informação fornecida pelo seu descritor. A interpretação envolve atribuição de 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.

Todas as tarefas das etapas descritas acima pressupõem a existência de um conhecimento sobre o problema a ser resolvido, armazenado em uma base de conhecimento, cujo tamanho e complexidade podem variar enormemente.

É importante destacar que a visualização dos resultados do processamento de imagens pode acontecer na saída de qualquer passo, e que também nem todas as aplicações de processamento de imagens requerem uma complexidade de interações como na Figura 2.2, em alguns casos, nem mesmo todos esses módulos são necessários (GONZALEZ e WOODS, 2000).

2.3.1 Equalização de histograma

O histograma de uma imagem digital com níveis de cinza no intervalo $[0, L-1]$ é uma função discreta $h(r_k) = n_k$, onde r_k é o k th nível de cinza e n_k é o número de pixels na imagem

de nível de cinza r_k (GONZALEZ e WOODS, 2000). O histograma de uma imagem indica o número de pixels que apresentam um determinado nível de cinza.

Uma das técnicas baseada em histograma, utilizada para realce de características, é a equalização de histograma. Essa técnica tem como finalidade obter um histograma uniforme através do espalhamento da distribuição dos níveis de cinza ao longo de toda a escala de contraste, aumentando dessa maneira, a detectabilidade de aspectos da imagem (relacionados a contraste) (ALMEIDA, 2010).

O histograma equalizado de uma imagem utilizando a Função de Distribuição Cumulativa (CDF) S_k pode ser obtido por:

$$S_k = T(r_k) = \sum_{j=0}^k \frac{n_j}{n} = \sum_{j=0}^k p_r(r_j) \quad (1)$$

onde $0 < r_k < 1$ (nível de cinza normalizado) e $k = 0, 1, 2, \dots, L-1$ (L é o número de níveis de cinza), n é o número total de pixels na imagem.

Como consequência dessa distribuição uniforme obtém-se uma imagem com melhor contraste, como é exemplificada na Figura 2.4.

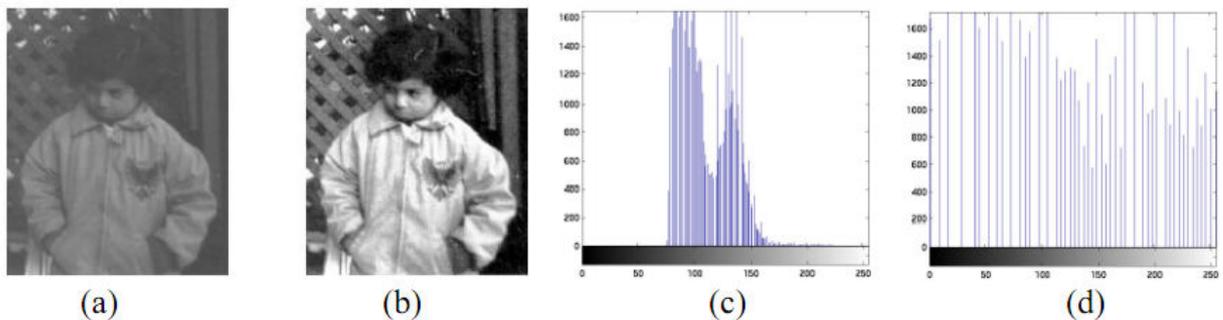


Figura 2.4 - Equalização de Histograma. (a) Imagem Original, (b) Imagem Equalizada, (c) Histograma Original e (d) Histograma Equalizado. Adaptado de (MATHWORKS, 2012).

Este trabalho utilizou a técnica de equalização de histograma da imagem para realçar características presentes nos nódulos, melhorando o desempenho das etapas posteriores da metodologia.

2.3.2 Quantização

A quantização de uma imagem digital consiste em quantizar o gamute⁴ de cores da imagem, o que acarreta na quantização da informação de cor de cada pixel da imagem. O processo de quantização transforma um conjunto de cores com M elementos em um conjunto de cores com N elementos, onde $M > N$ (GOMES e VELHO, 1994). Por exemplo, em uma imagem em tons de cinza, o conjunto de cores de entrada é subdividido em conjuntos menores denominados intervalos de quantização ou células de quantização. Todas as cores de uma célula de quantização assumem um mesmo valor de saída (um determinado nível de quantização), que corresponde ao valor representante dos elementos da célula (MARTINO, 2002). Na Figura 2.5(a) são exibidos três intervalos de quantização $[c_{i-1}, c_i]$, e em cada intervalo temos o nível de quantização associado q_i . Na Figura 2.5(b), é exibido o gráfico da função de quantização q , que é constante em cada intervalo de quantização.

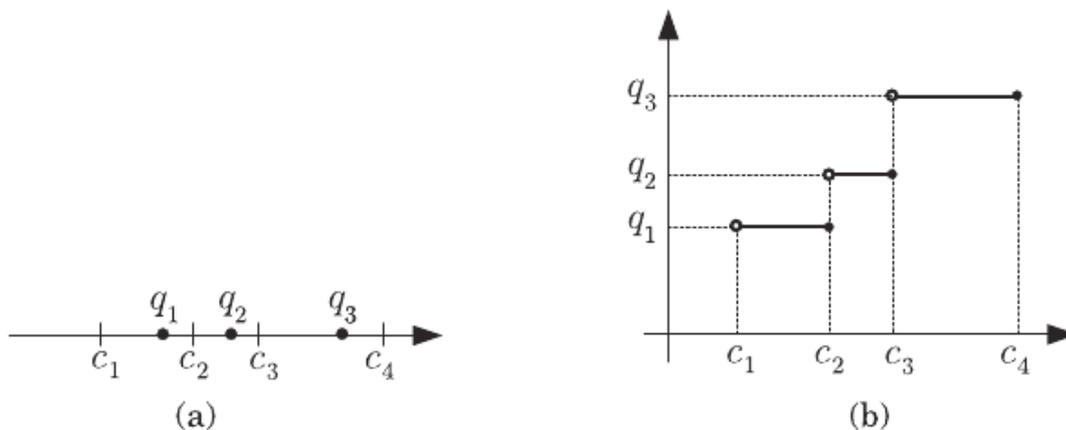


Figura 2.5 – Níveis de quantização e gráfico da função de quantização. Fonte: (GOMES e VELHO, 1994)

Para uma quantização unidimensional ou escalar (tons de cinza) a célula de quantização é um intervalo e pode-se variar apenas o seu comprimento. Na quantização multidimensional as células de quantização são regiões do espaço de cor que podem apresentar geometria bem mais complexa. (GOMES e VELHO, 1994).

Dentre as formas para determinar as células de quantização a escolha mais natural e simples à primeira vista, consiste em dividir o espaço de cor em células congruentes e em

⁴Corresponde ao espaço de cores possíveis em uma determinada imagem, por exemplo para uma imagem de 8 bits o gamute de cores corresponde aos valores dentre 0 e 255.

cada célula tomar o seu “centro” como sendo o nível de quantização associado. E no caso de quantização escalar com L níveis, as células de quantização são intervalos (c_{i-1}, c_i) de igual comprimento e em cada célula o valor de quantização é dado pela média (GOMES e VELHO, 1994), como apresentado em na Equação 2:

$$q_i = \frac{c_i + c_{i-1}}{2}, 1 \leq i \leq L \quad (2)$$

Esse método mais intuitivo e de fácil implementação é denominado quantização uniforme e será utilizado no trabalho 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 uma característica que visivelmente se descreve, porém, seu conceito é impreciso. Em (SONKA, HLAVAC e BOYLE, 1998) a textura é descrita como sendo um conjunto de elementos mutuamente relacionados formando blocos de textura que são agrupados em primitivas chamadas de *texels* (texture elements) e dependente da escala. Pode ser descrito pela sua tonalidade ou estrutura (MARTINS, 2007). A tonalidade está relacionada às propriedades do *pixel*, enquanto que a estrutura descreve a espacialidade das primitivas (SONKA, HLAVAC e BOYLE, 1998).

Diversas formas de obtenção de medidas de texturas têm sido buscadas para determinar a correlação espacial entre os padrões de textura (análise espacial entre pares de pixels em imagens 2D ou pares de *voxels* em imagens 3D) conforme (SILVA, SILVA, *et al.*, 2007, 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 governam a distribuição e a relação entre os níveis de cinza de uma região da imagem. Medidas estatísticas comuns incluem contraste, energia, entropia, correlação, homogeneidade, momento, que são obtidas da Matriz de co-ocorrência (HARALICK *et al.*, 1973).

Nas seções seguintes serão explanados os descritores de textura utilizados neste trabalho correspondentes a uma forma de abordagem estatística: os Índices de Diversidade de Shannon e de Simpson.

2.4.1 Índices de Diversidade

O termo diversidade, em ecologia, pode ser definido como a variedade e a variabilidade entre os organismos vivos e os complexos ecológicos em que ocorrem.

Uma medida de diversidade é um parâmetro extremamente reducionista que objetiva expressar toda a complexidade estrutural de uma comunidade ecológica através de um único número (SANTOS, 2009). Dentre a ampla gama de métodos disponíveis para medir a diversidade destacam-se, pelo amplo uso, Índices de Diversidade tais como os Índices de Shannon e Simpson que serão utilizados neste trabalho. Estes índices consistem de (ou confundem) dois componentes, riqueza de espécies e equabilidade e são usados há décadas por ecólogos, que examinaram suas habilidades em distinguir comunidades e suas propriedades matemáticas (MELO, 2008).

Outras alternativas de medida de diversidade incluem o uso de riqueza de espécies apenas, mas uma das vantagens da utilização de Índices de Diversidade é o fato de serem relativamente independentes do esforço amostral. A riqueza de espécies é bastante dependente do esforço amostral. Quanto mais coletamos, mais espécies encontramos. Os Índices de Diversidade, contudo, são pouco dependentes. Ou seja, com amostras relativamente pequenas podemos obter um valor de diversidade que mudará pouco conforme aumentamos o esforço amostral. Como consequência, podemos coletar pouco em cada comunidade a ser comparada e podemos comparar diretamente comunidades estudadas com diferentes esforços amostrais (MELO, 2008).

Neste trabalho, os Índices de Diversidade de Simpson e de Shannon são utilizados para medir a diversidade na distribuição dos *voxels* nas regiões de interesse de um volume de um nódulo. O objetivo é utilizar o valor obtido com os índices para caracterizar a textura das regiões segmentadas, para o reconhecimento de um padrão que possibilite ao classificador distinguir entre nódulos malignos e benignos.

2.4.1.1. Índice de Shannon

Originado da teoria da informação (SHANNON e WEAVER, 1949), o Índice de Shannon assume que os indivíduos são randomicamente amostrados a partir de uma comunidade infinitamente grande, e que todas as espécies estão representadas na amostra (MAGURRAN, 2004). O Índice de Shannon é calculado conforme a equação 3:

$$H' = -\sum_{i=1}^S p_i \ln p_i \quad (3)$$

Onde, p_i é a proporção de indivíduos pertencentes a espécie i , calculado como $p_i = n_i / N$, n_i é o número de indivíduos na espécie i e N é o número total de indivíduos na comunidade. S é o total de espécies.

Os valores obtidos através do Índice de Shannon (H') variam entre zero, onde existe apenas uma espécie, e o logaritmo de S , quando todas as espécies são representadas pelo mesmo número de indivíduos.

Para utilização nas imagens de TC, o Índice de Shannon foi calculado considerando os *voxels* no volume de um nódulo como os indivíduos da população, e as suas intensidades como às espécies.

2.4.1.2. Índice de Simpson

Proposto em 1949, por E. H. Simpson (SIMPSON, 1949), o Índice de Simpson é a medida de probabilidade de dois indivíduos selecionados aleatoriamente de uma comunidade infinitamente grande, pertencerem à mesma espécie (MAGURRAN, 2004). O Índice de Simpson é calculado conforme Equação 4:

$$D = \sum_{i=1}^j p_i^2 \quad (4)$$

Onde, $p_i = \frac{n_i}{N}$ é a probabilidade (p_i) para a ocorrência da espécie i , n representa a ocorrência de indivíduos da espécie i e N é o total de indivíduos da amostra.

Segundo (LYONS, DUNWORTH e TILBURY, 2009), para uma amostra finita o Índice de Simpson pode ser obtido ainda através da seguinte expressão:

$$D = \frac{\sum_{i=1}^j n_i(n_i - 1)}{N(N - 1)} \quad (5)$$

Os valores obtidos para o Índice de Simpson estão no intervalo entre 0 e 1, onde o valor 0 representa diversidade infinita e 1 representa que não há diversidade na amostra.

Para o contexto de processamento de imagens o índice de Simpson foi calculado de maneira análoga ao índice de Shannon, considerando os *voxels* em análise como sendo os

indivíduos da população, e as intensidades como às espécies. Assim consideramos que cada VOI utilizada possui uma distribuição onde n_i é o número total de *voxels* de uma tonalidade i presente na VOI e N é o número total de *voxels* da VOI.

2.5 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 (HAND, MANNILA e SMYTH, 2000; PAPPAS, 2002). Características irrelevantes ou redundantes podem confundir o algoritmo de aprendizagem, ajudando a esconder as distribuições de pequenos conjuntos de características realmente relevantes (KOLLER e SAHAMI, 1996).

A seleção de características pode ser vista como um processo de busca onde o algoritmo utilizado deve encontrar, de um conjunto de características, um subconjunto com a melhor eficiência no processo de classificação (OLIVEIRA, DUTRA e RENNÓ, 2005).

2.5.1 Análise Discriminante Linear *Stepwise*

A Análise Discriminante Linear é uma técnica que se tornou muito comum em aplicações de visão computacional. Essa técnica utiliza informações das classes associadas a cada padrão para extrair linearmente os atributos mais discriminantes através da computação de uma combinação linear de m variáveis quantitativas que mais eficientemente separam grupos de amostras em um espaço m -dimensional, fazendo com que a razão da variância intra-classes e inter-classes seja maximizada.

O problema é então reduzido a achar um vetor adequado (Equação 17). A idéia básica por trás da análise discriminante é determinar o quanto as classes são diferentes em relação à média de uma variável e depois usar essa variável para adequar um grupo para a nova amostra.

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x_r = \beta' x \quad (17)$$

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

A estimação simultânea envolve a computação da função discriminante, de modo que todas as variáveis independentes são consideradas juntas. Assim, a função discriminante é computada com base no conjunto inteiro de variáveis independentes, sem consideração do poder discriminatório de cada variável independente (HAIR, ANDERSON e BLACK, 2005).

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

Neste trabalho, é aplicada a Análise Discriminante Linear para determinar as variáveis que melhor discriminam os nódulos quanto à natureza (maligna ou benigna), utilizando o método *stepwise* para selecionar as variáveis independentes mais discriminantes das classes.

2.6 Reconhecimento de Padrões

Na natureza os padrões se manifestam de diversas maneiras como, por exemplo, sons, formas, imagens, cheiros e sabores. A todo instante os seres humanos, e também outros animais, percebem e interagem com estes padrões com extrema naturalidade. Exemplos disso são as habilidades que o ser humano tem de diferenciar o som do motor de um automóvel do som de uma música, ou ainda, a habilidade que um animal selvagem tem de distinguir uma presa de um predador. A naturalidade inerente a estas habilidades faz com que o ser humano sequer imagine as complexidades cognitivas que estão por trás delas. Complexidades que se tornam evidentes quando se tenta reproduzir as tais habilidades artificialmente em um computador, o que há muito desafia a comunidade científica interessada no assunto.

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 em visão computacional, análise sísmica, reconhecimento de voz, reconhecimento de faces, identificação de íris, identificação de digitais, reconhecimento de caracteres impressos e manuscritos, além de outras (NOQUEIRA, 2007).

A seguir será apresentada a formulação que fundamenta uma técnica utilizada para o reconhecimento de padrões, denominada Máquina de Vetores de Suporte (MVS).

2.6.1 Máquina de Vetores de Suporte

Introduzida por Vapnik (VAPNIK, 1988) a Máquina de Vetores de Suporte (MVS), é uma técnica de aprendizado supervisionado desenvolvida para classificação binária (duas classes).

Os algoritmos de treinamento das MVS possuem forte influência da teoria de otimização e de aprendizagem estatística. Em poucos anos, as MVS vêm demonstrando sua superioridade frente a outros classificadores em uma grande variedade de aplicações.

A idéia principal de uma Máquina de Vetores de Suporte é construir um hiperplano ótimo separando as classes. Por hiperplano entende-se uma superfície de separação de duas regiões em um espaço multidimensional, onde o número de dimensões possíveis pode ser, até, infinito. A Figura 2.6 mostra em duas dimensões, para melhor visualização, hiperplanos de separação entre duas classes linearmente separáveis. O hiperplano ótimo (linha mais escura), não somente separa as duas classes, mas mantém a maior distância possível com relação aos pontos da amostra.

Existem 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).

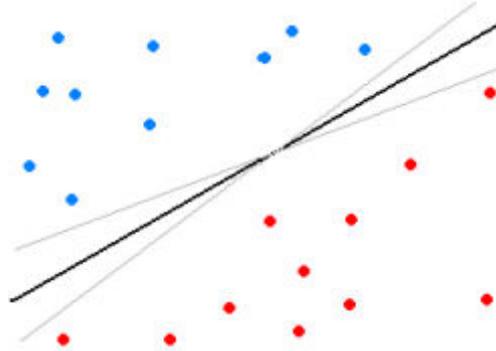


Figura 2.6 - Separação de duas classes através de hiperplanos.

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 $i = 1, 2, \dots, n$ o índice de cada ponto amostral. O objetivo da classificação é estimar a função $f: \mathbb{R}^n \rightarrow \{\pm 1\}$ que separe corretamente os exemplos de teste em classes distintas.

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

$$y_i((w \cdot x_i) + b) \geq 1 \quad (6)$$

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 6 ao minimizar a seguinte equação:

$$\Phi(w) = \frac{w^2}{2} \quad (7)$$

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 a inclusão de variáveis de folga, que permitem que as restrições presentes na equação 6 sejam quebradas.

O problema de otimização passa a ser então a minimização da Equação 8, de acordo com a restrição imposta pela equação 6. 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.

$$\Phi(w, \xi) = \frac{w^2}{2} + C \sum_{i=1}^N \xi_i \quad (8)$$

para

$$y_i((w \cdot x_i) + b) + \xi_i \geq 1 \quad (9)$$

Através da teoria dos multiplicadores de Lagrange, chega-se à Equação 10. O objetivo então passa a ser encontrar os multiplicadores de Lagrange a_i ótimos que satisfaçam a Equação 11 (CHAVES, 2006).

$$L(a) = \sum_{i=1}^N a_i - \frac{1}{2} \sum_{i,j=1}^N a_i a_j y_i y_j (x_i \cdot x_j) \quad (10)$$

$$\sum_{i=1}^N a_i y_i = 0, \quad 0 \leq a_i \leq C \quad (11)$$

Apenas os pontos onde a restrição dada pela Equação 6 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 2.7 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.

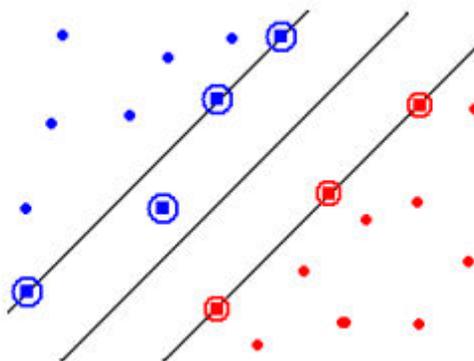


Figura 2.7 - Vetores de suporte (destacados por círculos).

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 ao invés 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 12 mostra o resultado da Equação 10 com a utilização de um núcleo K .

$$L(a) = \sum_{i=1}^N a_i - \frac{1}{2} \sum_{i,j=1}^N a_i a_j y_i y_j K(x_i, x_j) \quad (12)$$

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:

$$K(x_i, y_j) = \exp(-\gamma \|x_i - y_j\|^2) \quad (13)$$

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

2.7 Avaliação 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, mas 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 (S), Especificidade (E) e Acurácia (A) (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:

$$S = \frac{VP}{VP + FN} \quad (14)$$

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:

$$E = \frac{VN}{VN + FP} \quad (15)$$

A taxa de classificação correta (acurácia ou precisão) 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.

$$A = \frac{VP + VN}{VP + VN + FP + FN} \quad (16)$$

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

3 METODOLOGIA

Neste trabalho a metodologia proposta segue os passos da Figura 3.1. O primeiro é a aquisição das imagens que foram obtidas de exames de TC. O segundo é o pré-processamento das mesmas, através da equalização de histograma. O passo três é a segmentação do volume tri-dimensional do volume do nódulo. O quarto é a extração de características representativas dos nódulos, através do uso dos Índices de Shannon e de Simpson e, no quinto, as características são selecionadas, utilizando Análise Discriminante *stepwise*. Concluído o quinto passo, a classificação dos nódulos em benignos ou malignos é determinada pela MVS. O processo é finalizado com a validação dos resultados.



Figura 3.1- Estágios da Metodologia.

3.1 Aquisição das Imagens

Nesta seção, é descrita a aquisição das imagens para a extração de características. Obtiveram-se imagens de duas diferentes bases de dados. As bases de aquisição foram denominadas de Base I, para a origem através do Lung Image Database Consortium (LIDC) (ARMATO III, MCLENNAN, *et al.*, 2004) e de Base II para a origem através do Lung Image

Database Consortium e Image Database Resource Initiative (LIDC-IDRI) (ARMATO III, MCLENNAN, *et al.*, 2011). Nas próximas subseções aborda-se sobre cada uma delas.

3.1.1 Base I: LIDC

Neste trabalho foram obtidas imagens disponibilizadas pelo Instituto Nacional do Câncer dos EUA (NCI) criadas a partir de um repositório de imagens de câncer resultado da formação de um consórcio de instituições, conhecido como LIDC (Lung Image Database Consortium).

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

Na base do LIDC 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 entre 1 e 5), para aqueles nódulos pulmonares maiores que 3 mm, e apenas a informação sobre o centróide para aqueles inferiores a 3 mm. Dessa forma, de 84 exames disponíveis nesta base, só 58 possuem informação do contorno. Um contorno é formado por pares de pontos de coordenadas (x,y) que delimitam, em cada imagem do exame, a região onde o especialista identificou o nódulo.

O processo de anotação dos nódulos da base LIDC 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.

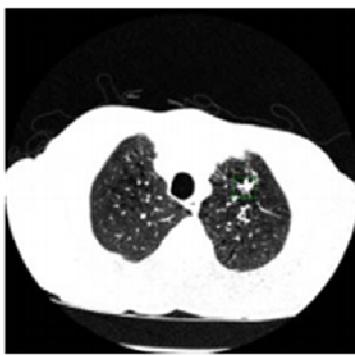
Não há imposição para que haja consenso, todos os nódulos indicados pela revisão dos radiologistas são apurados e gravados. Sendo assim, é possível ter diferentes diagnósticos para um mesmo nódulo. Considera-se então, neste trabalho, apenas uma instância por nódulo, objetivando minimizar o impacto da subjetividade nos exames. No entanto, não existe nenhuma indicação na anotação dos radiologistas (arquivo XML) sobre quais informações se referem ao mesmo nódulo. Para essa 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 3.2 ilustra o processo:



Figura 3.2 - Ilustração do resumo dos nódulos. Adaptado de (HORSTHEMKE, RAICU e FURST, 2008).

Na Figura 3.2, as linhas coloridas representam os contornos definidos pelos especialistas individualmente. O quadrado verde no centro refere-se ao centróide calculado para o contorno de mesma cor. Conforme cálculo, as coordenadas desse centróide se encontram nas áreas delimitadas por outros especialistas. Dessa forma, considera-se neste trabalho que tratam-se do mesmo nódulo e portanto só deve ser aceita uma instância, que será referente àquela marcação do nódulo que possuir maior área de contorno. No exemplo simplificado da Figura 3.2, a instância do nódulo que será utilizada será aquela representada pelo contorno de cor verde, pois atende a esse critério.

Após o cálculo de quais nódulos foram anotados por mais de um especialista, e de selecionar quais as instâncias correspondentes que serão utilizadas, é feito o resumo do diagnóstico quanto à malignidade ou benignidade. O diagnóstico já está presente para cada nódulo gravado na base, em uma escala subjetiva de cinco níveis que varia de altamente suspeito de benignidade a altamente suspeito de malignidade (representado no arquivo XML por números de 1 a 5), para o resumo então, utiliza-se as informações conforme levantamento da etapa anterior para que seja efetuado cálculo segundo apresentado em (JABON, RAICU e FURST, 2009), onde os valores das características pertencentes ao mesmo nódulo são reduzidos a um único valor através do cálculo da moda ou da mediana. Esse processo é ilustrado na Figura 3.3 destacando-se com o retângulo a característica de malignidade que será utilizada no trabalho.



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

Figura 3.3 - Ilustração do resumo do diagnóstico dos nódulos. Adaptado de (JABON, RAICU e FURST, 2009).

Em (JABON, RAICU e 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. Um aspecto a ressaltar é a proposta para gerar um resumo com menos desconformidade. 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, foram obtidos 73 nódulos sendo 47 benignos e 26 malignos.

3.1.2 Base II: LIDC-IDRI

Para a apuração dos resultados foi utilizada ainda a base de dados LIDC-IDRI (ARMATO III, MCLENNAN, *et al.*, 2011), disponibilizada recentemente como resultado de uma associação entre o Lung Image Database Consortium e a Image Database Resource Initiative. Essa parceria promoveu um aumento substancial da base LIDC resultando em um total de 1018 casos. Os exames disponibilizados por essa base apresentam as mesmas características da LIDC original, onde para cada um deles, existem as fatias correspondentes e o arquivo XML com as avaliações feitas por 4 especialistas.

Utilizando a mesma técnica para o resumo de nódulos proposta em (JABON, RAICU e FURST, 2009), foram considerados 1736 nódulos, sendo 1219 benignos e 517 malignos. Para diminuir o impacto do desbalanceamento, a base de dados foi dividida, totalizando 1024 nódulos: 517 nódulos malignos (escolhidos aleatoriamente) e 517 benignos para utilização neste trabalho.

3.2 Pré-Processamento

Durante a etapa de pré-processamento, realiza-se um realce de contraste da imagem através da equalização do histograma (Seção 2.2.1), com a intenção de melhorar a visualização de características presente nos nódulos.

Foi testada a equalização de histograma aplicada somente na área do nódulo e o uso das imagens não equalizadas. Porém a equalização de histograma considerando a imagem total foi a forma que possibilitou o melhor resultado.

3.3 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, a segmentação utilizada nesse trabalho segue o resumo apresentado na seção 3.1, onde somente a maior delimitação é escolhida para representar a instância dos nódulos descritos por até quatro especialistas.

3.4 Extração de Características

Após a segmentação dos nódulos, são calculadas algumas características que possam representá-los em uma das classes objetivadas nesse trabalho (maligna ou benigna). Nas subseções seguintes, são mostradas as formas de extração de características propostas.

3.4.1 Características de Textura

Para a extração de características aplicou-se duas forma de análises. Na primeira, as características foram obtidas considerando as áreas de interesse formadas por esferas concêntricas, como mostrado na Figura 3.4. Cada área de interesse representa um experimento do método. Na segunda análise, a área de interesse foi representada por anéis circulares, como na Figura 3.5. O objetivo com essas formas de extração é avaliar qual delas é a melhor maneira de utilizar a característica de textura, levando em consideração que a análise em círculos objetiva determinar diferenças no padrão de diversidade para nódulos malignos e benignos em cada região de estudo, da fronteira ao centro do nódulo. Na análise

em anéis, objetiva-se avaliar a diversidade nas bordas das regiões determinadas por duas esferas concêntricas.

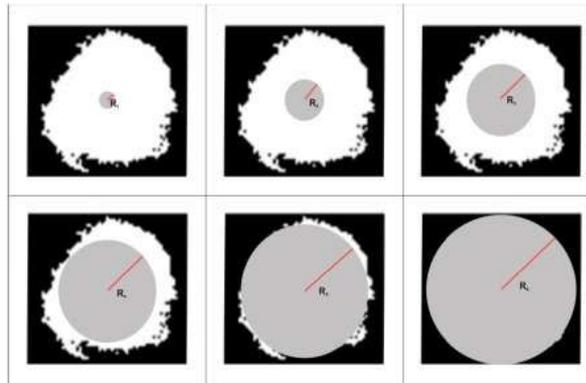


Figura 3.4 - Representação em 2D da análise em esferas concêntricas aplicada ao nódulo.

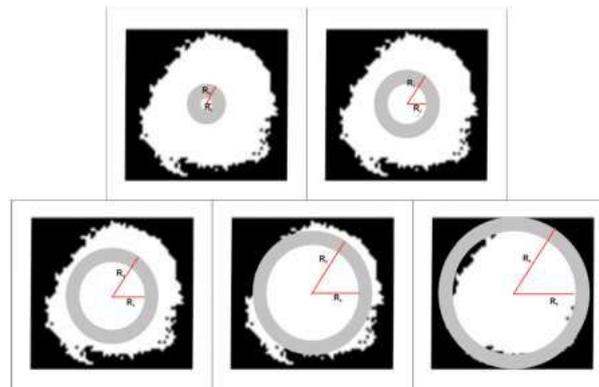


Figura 3.5 - Representação em 2D da análise em anéis esféricos aplicada ao nódulo.

O tamanho da esfera é determinado encontrando-se o ponto central de cada nódulo e então calculando a distância desse ponto central até o ponto mais distante de cada nódulo. Dessa forma, nós obtemos um raio R que representa a maior medida possível para a construção de uma esfera que circunscreva o nódulo, ou ainda, na análise em anéis, o máximo raio externo permitido. A partir do raio R , obtêm-se os outros valores de raios como $1/6R$, $1/3R$, $1/2R$, $2/3R$ e $5/6R$, que são representados por R_1 , R_2 , R_3 , R_4 , R_5 e R_6 (valor de R).

Encerrada essa etapa, os Índices de Shannon e de Simpson das Equações 3 e 5, respectivamente, são calculados.

Para a extração de características utilizou-se ainda três níveis de quantização (16, 12 e 8 bits) com o objetivo de realçar características dos nódulos presentes em diferentes escalas de tonalidade, melhorando a qualidade das informações para a extração de características com os Índices de Diversidade.

3.5 Seleção de características

Para realizar a seleção de características que melhor discriminam as classes maligna e benigna, foi utilizada a técnica Análise Discriminante *stepwise* (Seção 2.5.1). Para sua utilização é necessário um parâmetro F com valores utilizados como critério para entrada e remoção de variáveis. Estes valores foram sugeridos pelo software SPSS 17.0 (SPSS, 2010).

3.6 Classificação

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 radial da Equação 13 (Seção 2.6.1). Dessa maneira, é necessário estimar o melhor peso C e também o valor γ para a função radial. Os valores de C e γ são estimado com a base de treino através de uma busca exaustiva realizado pelo script desenvolvido em PYTHON: `grid.py` presente no pacote LIBSVM (CHANG e LIN, 2003).

3.7 Software e Hardware Utilizados

Para aplicação da metodologia foram desenvolvidos algoritmos na linguagem Java, através da IDE Netbeans 6.8. A biblioteca DCM4CHE (DCM4CHE, 2011), disponível gratuitamente na internet, foi utilizada para manipulação das imagens DICOM.

A biblioteca LIBSVM (CHANG e LIN, 2003) foi utilizada para o classificador MVS e também esta disponível gratuitamente na internet.

Para a seleção de características utilizando Análise Discriminante *stepwise*, foi utilizado o software Statistical Package for the Social Sciences (SPSS, 2010).

O computador utilizado para os testes foi um INTEL CORE 2 DUO 2,4 GHz, com 3 GB de memória RAM e HD com 250 GB.

4 TESTES E RESULTADOS

Neste capítulo serão apresentados e discutidos os resultados obtidos com a utilização da metodologia proposta neste trabalho para classificação de nódulos pulmonares de TC em maligno e benigno.

Para a realização dos testes, a partir da etapa de extração de características a base de amostras foi organizada em dois grupos para serem utilizados no classificador MVS: base de treinamento e base de teste.

Foram realizados 5 conjuntos treinamento/teste (Tr/Te) com 50% dos casos usados para treinamento e os outros 50% para teste (50/50). Cada conjunto foi selecionado de forma aleatória e submetido ao classificador.

As medidas de validação de resultados utilizadas são: sensibilidade, especificidade e acurácia. Para apresentação dessas medidas é computada a média dos valores obtidos em cada execução nos 5 conjuntos Tr/Te. Pretende-se, dessa forma, avaliar a metodologia de maneira menos dependente dos dados que são selecionados.

É importante ressaltar que, para avaliação dos resultados (médias), avalia-se também os valores das 5 classificações. Esses valores não podem exibir variação acentuada, pois isso indica forte dependência dos dados selecionados, e o que se procura é um método que seja robusto a esses fatores. Usa-se ainda como critério para avaliar os resultados, o equilíbrio entre as três medidas utilizadas (sensibilidade, especificidade e acurácia). No caso onde não há esse equilíbrio, busca-se aquele resultado que tem maior sensibilidade, pois entende-se que seja melhor informar sobre um diagnóstico de malignidade e despertar a partir desse diagnóstico uma atenção maior para o nódulo em estudo por parte do especialista.

Os produtos dos testes são dispostos segundo as abordagens apresentadas (anéis e esferas) após a seleção de características, utilizando índice de diversidade de Shannon e Simpson separadamente e em conjunto, para cada base de dados (LIDC e LIDC-IDRI). Na apresentação dos resultados são exibidos os valores das médias e o desvio padrão em relação às 5 classificações efetuadas. 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 abreviaturas A1, A2, A3, A4 e A5 para os anéis e R1, R2, R3, R4, R5 e R6 para as esferas.

4.1 Base I: LIDC

Nesta seção são apresentados os testes utilizando a base LIDC, com 73 nódulos, sendo 26 malignos e 47 benignos.

4.1.1 Testes com o Índice de Shannon

Nas subseções seguintes são apresentados os resultados obtidos nas abordagens propostas na seção 3.4.1 para extração de características, utilizando o Índice de Shannon com a técnica de seleção de características.

4.1.1.1. Abordagem em anéis

Utilizando a abordagem em anéis (Figura 3.5), é gerado um total de 15 variáveis (5 anéis em 3 diferentes quantizações).

Após a etapa de seleção de características, apenas 2 foram selecionadas. As variáveis selecionadas referem-se ao anel A1 em dois diferentes níveis de quantização: Q12 e Q16. Essas variáveis foram então utilizadas no classificador MVS e as médias dos valores das medidas avaliativas após a etapa de classificação são apresentados na Tabela 1.

Tabela 1 - Resultados dos testes utilizando o Índice de Shannon, em anéis, extraído da Base I.

Variáveis selecionadas: A1 de Q12 e A1 de Q16			
Sensibilidade (%)	Especificidade(%)	Acurácia(%)	
77,97	92,17	86,67	Média
19,18	4,95	5,34	Desv. Padrão

De acordo com a Tabela 1, utilizando-se as 2 variáveis selecionadas correspondentes ao anel de raio mais interno com as quantizações com maiores níveis de bits por *voxel*, verifica-se que há um desbalanceamento entre sensibilidade e especificidade (essa maior). Além disso, a sensibilidade possui desvio padrão alto. Atribui-se essa dificuldade e alternância na caracterização de nódulos malignos ao desbalanceamento da base com maioria de nódulos sendo de natureza benigna.

4.1.1.2. Abordagem em esferas

Utilizando o Índice de Shannon na abordagem em esferas (Figura 3.4), são geradas 18 variáveis (6 esferas em 3 diferentes quantizações).

Posteriormente a etapa de seleção de características, apenas 1 foi selecionada, correspondente a esfera R1 de Q16. Essa variável então, seguindo o fluxo da metodologia, foi utilizada no classificador MVS. Na tabela 2 estão os resultados oriundos à conclusão da fase de classificação.

Tabela 2 - Resultados dos testes utilizando o Índice de Shannon, em esferas, extraído da Base I

Variável selecionada: R1 de Q16			
Sensibilidade (%)	Especificidade(%)	Acurácia(%)	
85,64	97,89	92,78	Média
5,63	2,88	3,16	Desv. Padrão

Utilizando a variável selecionada referente à esfera de raio mais interno e quantização com maior nível de bits por *voxel*, o resultado se mostrou eficiente, como mostra a Tabela 2. Os resultados exibem desvios padrão baixos em relação à média, para todas as medidas avaliativas. E apesar de haver apenas uma característica selecionada, ela se mostrou com bom poder discriminatório. O desbalanceamento entre as medidas de sensibilidade e especificidade novamente é atribuído ao desbalanceamento da base de dados que fornece maior quantidade de amostras de natureza benigna, facilitando assim a predição para a classe benigna.

4.1.2 Testes com o Índice de Simpson

Nesta seção são apresentados os resultados para o Índice de Simpson nas mesmas configurações da apuração dos resultados para o Índice de Shannon.

4.1.2.1. Testes utilizando abordagem em anéis

Aqui encontram-se descritos os resultados obtidos com a abordagem em anéis para o Índice de Simpson. Dentre as 15 características disponíveis, apenas 3 foram selecionadas após aplicação da Análise Discriminante *stepwise*. As variáveis selecionadas foram A1 de Q8, A4 e A5 de Q16. Os resultados oriundos à conclusão da etapa de classificação, utilizando essas variáveis, são apresentados na Tabela 3.

Tabela 3 - Resultados dos testes utilizando o Índice de Simpson, em anéis, extraído da Base I.

Variáveis selecionadas: A1 de Q8, A4 e A5 de Q16			
Sensibilidade (%)	Especificidade(%)	Acurácia(%)	
78,65	81,45	80,00	Média
18,45	8,51	8,65	Desv. Padrão

Na Tabela 3, as medidas de sensibilidade, especificidade e acurácia apresentam equilíbrio, no entanto observa-se que a medida de sensibilidade apresenta desvio padrão mais alto, assim como foi observado para o Índice de Shannon na mesma abordagem. As demais medidas apresentam desvios padrão moderados.

Para o índice de Simpson, mais variáveis foram selecionadas, integrando tanto anéis mais internos quanto os mais externos, e as quantizações com poucos e muitos bits por voxel. No entanto, os resultados são inferiores aos obtidos pelo Índice de Shannon para a mesma abordagem.

4.1.2.2. Testes utilizando abordagem em esferas

Nesta seção estão descritos os resultados obtidos com a abordagem em esferas. Na Tabela 4 é possível observar que esses resultados foram obtidos utilizando-se apenas 1 característica, referente à esfera R6 de Q16, das 18 características disponíveis para essa configuração.

Tabela 4 - Resultados dos testes utilizando o Índice de Simpson, em esferas, extraído da Base I.

Variável selecionada: R6 de Q16			
Sensibilidade (%)	Especificidade(%)	Acurácia(%)	
52,85%	80,51%	65,56%	Média
35,55	17,12	15,54	Desv. Padrão

Na Tabela 4 pode-se observar desvios padrão acentuados em todas as medidas, tornando assim os resultados pouco significativos.

Assim como para o Índice de Shannon, nesta configuração, para o Índice de Simpson também apenas uma variável foi selecionada. Entretanto, foi escolhida uma esfera de maior raio e os resultados não foram satisfatórios.

4.1.3 Testes com o Índice de Shannon e de Simpson

Nesta seção são exibidos os resultados para o Índice de Shannon e Simpson utilizados conjuntamente. Contudo, não são apresentados os resultados na abordagem em anéis e em anéis e esferas conjuntamente, uma vez que as variáveis selecionadas e os resultados de classificação já estão presentes na tabela 1 e na Tabela 5, respectivamente.

Tabela 5 - Resultados dos testes utilizando os Índices de Shannon e Simpson, em esferas, extraídos da Base I.

Variáveis selecionadas: R1 Simpson de Q8, R2 Shannon de Q8, R1 Shannon de Q16			
Sensibilidade (%)	Especificidade(%)	Acurácia(%)	
84.64	94.36	90.56	Média
8,76	4,18	3,17	Desv. Padrão

A Tabela 5 exibe os resultados com as 3 variáveis selecionadas, contemplando ambos índices de diversidade das quantizações Q8 e Q16. O desvio padrão é moderado para a medida de sensibilidade e baixo para as demais medidas. O resultado é relativamente próximo ao obtido pela análise em esferas utilizando somente o Índice de Shannon, tanto em relação às médias quanto aos desvios padrão.

Para a abordagem em anéis, somente variáveis do Índice de Shannon foram selecionadas. Portanto, o teste não foi repetido (Tabela 1). Para a abordagem em anéis e esferas somente variáveis da abordagem em esferas foram selecionadas e assim o teste também não foi repetido (Tabela 5). Observou-se que para a base de dados utilizada, o índice que mais se destaca é o Índice de Shannon, bem como a utilização de esferas na forma de abordagem para a extração de características.

4.2 Base II: LIDC-IDRI

Nesta seção são descritos os resultados para a Base II: LIDC-IDRI, que possui um número significativamente maior de casos para utilização no experimento (1034 nódulos balanceados) em relação à Base I. As configurações de testes são as mesmas aplicadas anteriormente: Índice de Shannon e Simpson em anéis e esferas.

4.2.1 Testes com o Índice de Shannon

A seguir serão apresentados os resultados obtidos nas abordagens anéis e esferas, utilizando o Índice de Shannon com seleção de características.

4.2.1.1. Abordagem em anéis

No teste para a abordagem em anéis, foram selecionadas 7 características: A1 de Q8, A3 de Q12, e os 5 anéis de Q16. Os resultados são apresentados na Tabela 6.

Tabela 6 - Resultados dos testes utilizando o Índice de Shannon, em anéis, extraído da Base II.

Variáveis selecionadas: A1 de Q8, A3 de Q12, A1, A2, A3, A4 e A5 de Q16			
Sensibilidade (%)	Especificidade(%)	Acurácia(%)	
82,63	85,19	83,91	Média
3,73	1,90	1,66	Desv. Padrão

Conforme Tabela 6, as medidas de sensibilidade, especificidade e acurácia, apresentam resultados razoáveis, com equilíbrio entre os valores e baixos desvios padrão. As variáveis selecionadas contemplam os mais diferentes raios de anéis e a quantização que mais contribuiu na análise foi Q16.

4.2.1.2. Abordagem em esferas

Para a abordagem em esferas no Índice de Shannon foram selecionadas 5 variáveis: R1 de Q8, R6 de Q12, R1, R5 e R6 de Q16.

Tabela 7 - Resultados dos testes utilizando o Índice de Shannon, em esferas, extraído da Base II.

Variáveis selecionadas: R1 de Q8, R6 de Q12, R1, R5 e R6 de Q16			
Sensibilidade (%)	Especificidade(%)	Acurácia(%)	
81,87	85,14	83,52	Média
1,30	2,05	0,68	Desv. Padrão

Conforme Tabela 7, para as variáveis selecionadas na abordagem em esferas os resultados são ligeiramente inferiores aos obtidos pela abordagem em anéis, no entanto possuem desvios padrão um pouco mais baixos. Novamente, a quantização que mais contribuiu em número de características selecionadas foi Q16.

4.2.2 Testes com o Índice de Simpson

Nesta seção são apresentados os resultados para o Índice de Simpson nas abordagens em anéis e esferas.

4.2.2.1. Testes utilizando abordagem em anéis

Para os resultados obtidos com a abordagem em anéis utilizando Índice de Simpson foram utilizadas 4 características selecionadas: A1, A4 de Q8, A1 e A5 de Q12. Os resultados da classificação para essas variáveis são exibidos na Tabela 8.

Tabela 8 - Resultados dos testes utilizando o Índice de Simpson, em anéis, extraído da Base II.

Variáveis selecionadas: A1, A4 de Q8, A1 e A5 de Q12			
Sensibilidade (%)	Especificidade(%)	Acurácia(%)	
75,32	55,32	65,26	Média
13,83	7,76	3,85	Desv. Padrão

Conforme Tabela 8, os resultados apurados tiveram valores baixos, e embora a sensibilidade apresente uma média maior em relação à especificidade, teve um desvio padrão mais acentuado. Diante do exposto, o Índice de Simpson isoladamente se mostrou inferior ao Índice de Shannon para essa abordagem.

4.2.2.2. Testes utilizando abordagem em esferas

Nesta seção estão os resultados obtidos com a abordagem em esferas na Tabela 9. Para esses resultados foram utilizadas 7 variáveis: R1, R2, R3, R5 de Q8, R1, R4 de Q12 e R6 de Q16.

Tabela 9 - Resultados dos testes utilizando o Índice de Simpson, em esferas, extraído da Base II.

Variáveis selecionadas: R1, R2, R3, R5 de Q8, R1, R4 de Q12 e R6 de Q16			
Sensibilidade (%)	Especificidade(%)	Acurácia(%)	
81,97	67,40	74,66	Média
7,37	6,54	1,90	Desv. Padrão

Nessa análise, conforme Tabela 9, os resultados foram superiores a abordagem em anéis, mas ainda inferiores aos resultados obtidos com o Índice de Shannon. O desvio padrão é mediano para sensibilidade e especificidade.

4.2.3 Testes com o Índice de Shannon e Simpson

Nesta seção são apresentados os resultados para os Índices de Shannon e de Simpson utilizados conjuntamente nas abordagens em anéis e esferas.

4.2.3.1. Testes utilizando abordagem em anéis

Para os resultados obtidos com essa abordagem, empregando ambos os índices de diversidade, foram utilizadas 9 características selecionadas de um total de 30 (5 anéis para cada índice de diversidade em 3 quantizações), são elas: A1 Shannon de Q8, A1, A4 Simpson de Q8, A3 Shannon de Q12, A2 Simpson de Q12, A1, A2, A3 e A5 Shannon de Q16. Os resultados da classificação estão na Tabela 10.

Tabela 10 - Resultados dos testes utilizando os Índices de Shannon e Simpson, em anéis, extraídos da Base II

Variáveis selecionadas: A1 Shannon de Q8, A1, A4 Simpson de Q16, A3 Shannon de Q12, A2 Simpson de Q12, A1, A2, A3 e A5 Shannon de Q16			
Sensibilidade (%)	Especificidade(%)	Acurácia(%)	
82,95	84,58	83,75	Média
3,25	2,50	2,20	Desv. Padrão

Os valores apresentados na Tabela 10 demonstram uma capacidade de predição razoável, com valores equilibrados entre as médias e com desvios padrão baixos. As variáveis selecionadas contemplam todos os anéis e todas as quantizações, sendo que a quantização Q16 foi a que mais contribuiu com características selecionadas.

4.2.3.2. Testes utilizando abordagem em esferas

Nesta seção estão os resultados obtidos com a abordagem em esferas. Eles foram obtidos utilizando-se 10 variáveis: R2 Shannon de Q8, R2 Simpson de Q8, R1, R2, R5, R6 Shannon de Q12, R1, R4 Simpson de Q12, R3 e R6 Shannon em Q16. Os resultados da classificação estão na tabela 11.

Tabela 11 - Resultados dos testes utilizando os Índices de Shannon e Simpson, em esferas, extraídos da Base II.

Variáveis selecionadas: R2 Shannon de Q8, R2 Simpson de Q8, R1, R2, R5, R6 Shannon de Q12, R1, R4 Simpson de Q12, R3 e R6 Shannon de Q16			
Sensibilidade (%)	Especificidade(%)	Acurácia(%)	
81,72	85,50	83,64	Média
3,81	2,36	1,87	Desv. Padrão

Conforme Tabela 11, os resultados foram bem semelhantes aos obtidos com a abordagem em anéis (Tabela 10), tanto nos valores de média como nos desvios padrão. As variáveis selecionadas utilizaram todos os raios de esferas e todas as quantizações.

4.2.3.1 Testes utilizando abordagem em anéis e esferas

Nesta seção estão os resultados obtidos com a abordagem em anéis e esferas conjuntamente. Esses resultados foram obtidos utilizando-se 10 variáveis: A4 Simpson de Q8, R1 Simpson de Q12, A3 Simpson de Q12, R1, R5 Shannon de Q12, A4 Shannon de Q12, R2 R3, R6 Shannon de Q16 e A1 Shannon de Q16. Os resultados da classificação utilizando a MVS estão presentes na Tabela 12.

Tabela 12 - Resultados dos testes utilizando os Índices de Shannon e Simpson, em anéis e esferas, extraídos da Base II.

Variáveis selecionadas: A4 Simpson Q8, R1 Simpson de Q12, A3 Simpson de Q12, R1, R5 Shannon de Q12, A4 Shannon de Q12, R2, R3, R6 Shannon de Q16 e A1 Shannon Q16

Sensibilidade (%)	Especificidade(%)	Acurácia(%)	
82,33	87,13	84,76	Média
2,10	1,84	1,30	Desv. Padrão

Na Tabela 12 as medidas apontam também valores bastante razoáveis para a discriminação entre nódulos malignos e benignos, com baixos desvios padrão. As variáveis selecionadas contemplam diferentes raios de anéis e esferas, em diferentes quantizações, nos dois Índices de Diversidade. Os valores obtidos nessa configuração são muito próximos aos obtidos em outras abordagens, assim como observado para a Base I. Entretanto, novamente o Índice de Simpson parece não contribuir muito nos resultados quando utilizado em conjunto com o Índice de Shannon, e apresenta resultados também pouco relevantes individualmente.

4.3 Avaliação Final

Neste item pretende-se sintetizar os resultados obtidos com a aplicação da metodologia proposta, de forma a melhorar a compreensão de uma maneira geral do que foi apurado nos testes.

Para os resultados da Base I foi possível observar em boa parte do experimento que a medida de especificidade apresentou valores mais altos do que a medida de sensibilidade.

Atribui-se a isso o fato da maioria dos nódulos considerados para essa base serem de natureza benigna. Desta forma, para a classe maligna o classificador possui uma quantidade menor de amostras para o aprendizado. Ainda na Base I, o Índice de Shannon, assim como a quantização Q16 e a abordagem em esferas, destacaram-se isoladamente.

Utilizando a Base II com um número muito maior de casos, observou-se que alguns comportamentos se mantiveram, como por exemplo, a superioridade do Índice de Shannon individualmente nos resultados em relação ao Índice de Simpson. Entretanto, para essa base, configurações utilizando diferentes raios de anéis e esferas e diferentes níveis de quantização apresentaram também resultados razoáveis e mesmo sendo inferiores em relação ao que foi obtido com a Base I, apresentaram menor desvio padrão.

Com o balanceamento da Base II em relação à quantidade de nódulos malignos e benignos, foi possível observar também que os resultados apresentaram equilíbrio nas medidas avaliativas, reafirmando a sugestão inicial sobre a influência do desbalanceamento entre as classes para os resultados da Base I.

Mesmo que a Base II tenha apresentado um resultado sutilmente superior utilizando uma configuração com o Índice de Shannon e Simpson em conjunto, atenta-se sobre a própria aleatoriedade dos dados selecionados para classificação, que pode ocasionar algumas variações nos resultados. Portanto, conclui-se que na metodologia e bases utilizadas, o Índice de Shannon se mostrou suficiente, apresentando resultados melhores, já o Índice de Simpson não ofereceu uma contribuição significativa.

A Tabela 13 resume os melhores resultados nos índices de diversidade aplicados após a etapa de seleção de características.

Tabela 13 - Resumo dos melhores resultados obtidos pela metodologia em cada índice de diversidade isoladamente e em conjunto, para cada base de dados.

MELHORES RESULTADOS					
Base I					
Índice(s) de Diversidade	S(%)	E(%)	A(%)	Abordagem Utilizada	Quantizações Utilizadas
Shannon	85,64	97,89	92,78	Esferas	Q16
Simpson	78,65	81,45	80,00	Anéis	Q8 e Q12
Shannon e Simpson	84,64	94,36	90,56	Esferas	Q8 e Q16
Base II					
Shannon	82,63	85,19	83,91	Anéis	Q8, Q12 e Q16
Simpson	81,97	67,40	74,66	Esferas	Q8, Q12 e Q16
Shannon e Simpson	82,95	84,58	83,75	Anéis	Q8, Q12 e Q16

Analisando os trabalhos apresentados na literatura, podemos observar que a metodologia proposta consegue resultados comparáveis aos melhores já publicados, conforme se observa na Tabela 14.

Comparando-se os resultados para a base LIDC, é possível observar que houve um aumento da precisão na caracterização de nódulos pulmonares com a metodologia apresentada. E embora que ainda alguns valores, para algumas medidas, sejam inferiores na comparação com outros trabalhos, é preciso observar aspectos como o fato da base utilizada em (EL-BAZ, NITZKEN, et al., 2010; EL-BAZ, GIMEL'FARB, et al., 2011) apresentar 109 nódulos e a base HUPE apresentar apenas 39 nódulos, sendo que se observa na literatura, como no próprio trabalho de (SILVA, 2009), que aumentando-se o número de amostras as medidas de sensibilidade, especificidade e acurácia tendem a adquirir valores mais baixos. Levando-se em conta essas observações é possível destacar que os experimentos realizados no presente trabalho na tarefa de classificação de nódulos pulmonares em maligno e benigno se mostram bastante efetivos e promissores, encorajando estudos mais profundos, considerando, inclusive, a utilização em conjunto com as outras metodologias existentes.

Tabela 14 – Comparação dos resultados entre trabalhos relacionados.

Trabalhos	Base utilizada	S(%)	E(%)	A(%)
(SILVA, 2009)	HUPE	90,00	96,67	95,00
	LIDC	89,70	89,70	89,70
(EL-BAZ, NITZKEN, et al., 2010)	Proprietária	92,15	94,82	94,49
(EL-BAZ, GIMEL'FARB, et al., 2011)	Proprietária	92,00	100,00	96,36
METODOLOGIA PROPOSTA	LIDC	85,64	97,89	92,78
	LIDC-IDRI	82,95	84,58	83,75

Finalmente é importante salientar também que, para uma comparação completamente justa das metodologias citadas, seria necessário utilizar as mesmas imagens em todos os trabalhos. Além disso, deveria haver alguns parâmetros padrão, tais como resolução, bits por voxel, protocolo, etc. 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.

5 CONCLUSÃO

Neste trabalho, foi apresentada a utilização dos Índices de Diversidade de Shannon e de Simpson submetidos ao classificador Máquina de Vetores de Suporte para o diagnóstico de nódulos pulmonares em imagens de TC.

Os melhores resultados da metodologia foram obtidos utilizando o Índice de Shannon isoladamente. Este índice demonstrou ser superior e suficiente para a discriminação dos nódulos em relação ao índice de Simpson, nas bases de dados e abordagens utilizadas. A utilização da abordagem em esferas, raios mais internos e das imagens originais se destacou para a Base I, entretanto observa-se que diferentes raios de anéis e esferas e diferentes níveis de quantização também contribuíram para alguns dos melhores resultados obtidos no trabalho. Isso demonstra que a técnica de quantização e a utilização de ambas abordagens foram relevantes para a extração dos Índices de Diversidade.

É importante destacar que uma das dificuldades do trabalho foi em relação às próprias bases de dados utilizadas, pelo fato de possuírem diagnóstico realizado de forma subjetiva, com anotação feita por diferentes radiologistas. Contudo, observa-se grande reconhecimento acadêmico dessas bases e vários trabalhos (HORSTHEMKE, RAICU e FURST, 2008; VARUTBANGKUL, MITROVIC, *et al.*, 2008; JABON, RAICU e FURST, 2009) que apresentam estudos para o melhor tratamento das informações pertencentes a elas. Desta forma, reafirma-se a relevância dos resultados obtidos e da metodologia apresentada para a literatura. No entanto ainda assim, persistiram em existir aspectos que durante o desenvolvimento do trabalho não puderam ser concluídos ou inclusos. A investigação dos mesmos poderá promover o melhoramento da metodologia, são eles:

- Avaliação das medidas de geometria presentes em (Souza, 2009) com os índices de diversidade apresentados, para a caracterização de nódulos pulmonares;
- Avaliação de outros índices de diversidade na tarefa de classificação dos nódulos;
- Melhoramento da segmentação dos nódulos do resumo;
- Investigação de outras técnicas de seleção de características, como Análise de Componente Principais (PCA) e Algoritmos Genéticos (como o GA-SVM).
- Averiguação e adaptação da metodologia em outras áreas de estudo como na classificação de massa e não massa utilizando imagens mamográficas.

REFERÊNCIAS

- ALMEIDA, J. D. S. **Metodologia Computacional para Detecção automática de Estrabismo em imagens digitais através do teste de Hirschberg**. Dissertação (Mestrado). Universidade Federal do Maranhão. São Luís, p. 100. 2010.
- ARMATO III, S. G. et al. The Lung Image Database Consortium Research Group: Lung Image Database Consortium: Developing a resource for the medical imaging research community. **Radiology**, v. 232, p. 739–748, 2004.
- 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,p. 915–931, 2011.
- AZEVEDO-MARQUES, P. M. Diagnóstico auxiliado por computador na radiologia. **Radiologia Brasileira**, v. 34, p. 285-293, 2001.
- BLAND, M. **An Introduction to Medical Statistics**. Oxford: Oxford University Press, v. 132, 2000. 405 p.
- 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.
- CAMPOS, V. O. **Segmentação Multicritério para Detecção de Nódulos Pulmonares em Imagens de Tomografia Computadorizada**. Tese (Doutorado). Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro. p. 110. 2009.
- 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. G. Nódulo pulmonar. **Revista Brasileira de Medicina**, v. 68, p. 16-21, 2011.

CHAVEZ, A. C. F. **Extração de regras Fuzzy para Máquinas de vetor de suporte (SVM) para classificação em múltiplas classes.** Tese (Doutorado). Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro. 2006.

DCM4CHE , 2011. Disponível em: <<http://www.dcm4che.org/>>.

EL-BAZ, A. et al. Appearance analysis for diagnosing malignant lung nodules. **Proc. of IEEE International Symposium on Biomedical Imaging: From Nano to Macro.** Rotterdam, p. 193-196. 2010.

EL-BAZ, A. et al. A novel shape-based diagnostic approach for early diagnosis of lung nodules. **Proc. of IEEE International Symposium on Biomedical Imaging: From Nano to Macro.** Chicago, p. 137-140, 2011.

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

GOMES, J.; VELHO, L. **Computação Gráfica: Imagem.** Rio de Janeiro: IMPA/SBM, 1994. 424 p.

GONZALEZ, R. C.; WOODS, R. E. **Processamento de Imagens Digitais.** São Paulo: Edgard Blücher, 2000. 509 p.

HAIR, J. F. J.; ANDERSON, R. E. T. R. L.; BLACK, W. C. **Análise Multivariada de Dados.** Porto Alegre: Bookman, 2005. 593 p

HAND, D. J.; MANNILA, H.; SMYTH, P. **Principles of Data Mining.** Cambridge: A Bradford Book, 2000. 546 p.

HAYKIN, S. **Redes Neurais: Princípios e Práticas.** Porto Alegre: Bookman, 2001. 236p.

HORSTHEMKE, W. H.; RAICU, D. S.; FURST, J. D. Evaluation Challenges for Bridging Semantic Gap: Shape Disagreements on Pulmonary Nodules in the Lung Image Database Consortium. **International Journal of Healthcare Information Systems and Informatics (IJHISI) Special Edition on Content-based Medical Image Retrieval**, v. 4, p 17-33, 2009.

INCA - INSTITUTO NACIONAL DE CÂNCER. **O que causa o câncer?**, 2011. Disponível em: <http://www.inca.gov.br/conteudo_view.asp?id=81>. Acesso em: 01 dez. 2011.

INCA - INSTITUTO NACIONAL DE CÂNCER. **Tipos de Câncer: Pulmão**, 2012. Disponível em: <<http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/pulmao>>. Acesso em: 05 jan. 2012.

JABON, S. A.; RAICU, D. S.; FURST, J. D. Content-based versus Semantic-based Retrieval: A LIDC Case Study. **Progress in biomedical optics and imaging**, v. 10, p. 72631L.1-72631L.8, 2009.

KOLLER, D.; SAHAMI, M. **Toward Optimal Feature Selection**. Proc. Of 13th International Conference in machine Learning. [S.l.]:[s.n.]. 1996. p. 284-292.

KUMAR, S. A. et al. **Robust and Automated Lung Nodule Diagnosis from CT images Based on Fuzzy Systems**. International Conference on Process Automation, Control and Computing, Coimbatore: [s.n.]. 2011. p 954-959.

LEE, S. L. A.; KOUZAN, A. Z.; NASIERDING, G. **Pulmonary Nodule Classification Aided by Clustering**. IEEE International Conference on Systems, Man, and Cybernetics. San Antonio:[s.n.]. 2009. p. 906-911.

LIU, Y.; YANG, J.; ZHAO, D. **Computer Aided Detection of Lung Nodules Based on Voxel Analysis utilizing Support Vector**. BioMedical Information Engineering, 2009. FBIE 2009. International Conference on Future. Sanya:[s.n.]. 2009. p.90-93.

LYONS, D. J.; DUNWORTH, P. M.; TILBURY, D. W. **Simpson's Diversity Index**, 2009. Disponível em: <<http://www.countrysideinfo.co.uk/simpsons.htm>>. Acesso em: 13 jan. 2012.

MAGURRAN, A. E. **Measuring Biological Diversity**. Oxford: Blackwell Science, 2004. 256 p.

MARTINO, J. M. D. **Sistemas de Informações Gráficas**. [S.l.]. 2002.

MARTINS, L. O. **Detecção de massas em imagens mamográficas através do algoritmo Growing Neural Gas e da Função K de Ripley**. Dissertação (Mestrado). Universidade Federal do Maranhão. São Luís. p. 105. 2007.

MATHWORKS. **Adjusting Pixel Intensity Values**, 2012. Disponível em: <<http://www.mathworks.com/help/toolbox/images/fl1-14011.html>>. Acesso em: 20 jan. 2012.

MELO, A. S. O que ganhamos ‘confundindo’ riqueza de espécies e equabilidade em um índice de diversidade? **Biota Neotrop**, v. 8, 2008.

NOGUEIRA, A. et al. Um Overview Sobre Reconhecimento de Padrões, 2006. Disponível em: <http://www.aedb.br/seget/artigos06/688_SEGET_Um_Overview_Sobre_Reconhecimento_de_Padr_es.pdf>. Acesso em: 20 Janeiro 2012.

OLIVEIRA, J. A.; DUTRA, L. V.; RENNÓ, C. D. **Aplicações de métodos de extração e seleção de atributos para classificação de regiões**. Anais XII Simpósio Brasileiro de Sensoriamento Remoto. Goiânia: [s.n.]. 2005. p. 4201-4208.

PAPPA, G. L. **Seleção de atributos utilizando algoritmos genéticos multiobjetivos**. Dissertação (Mestrado). Pontifícia Universidade Católica do Paraná. Curitiba, p. 85. 2002.

SANTOS, A. M. **Detecção de nódulos pulmonares pequenos usando Modelo de Mistura Gaussiana e Matriz Hessiana**. Dissertação (Mestrado). Universidade Federal do Maranhão. São Luís, p. 81. 2011.

SANTOS, V. K. **Uma generalização da distribuição do índice de diversidade generalizada por Good com aplicação em Ciências Agrárias**. Dissertação (Mestrado). Universidade Federal Rural de Pernambuco. Recife, p. 57. 2009.

SHANNON, C. E.; WEAVER, W. **The mathematical theory of communication**. [S.l.]: University of Illinois Press, 1949.

SILVA, A. C.; CARVALHO, P. C. P. **Sistema de análise de nódulo pulmonar**. II Workshop de Informática aplicado a Saúde. Universidade de Itajaí. Itajaí. 2008.

SILVA, C. A. **Caracterização de nódulos pulmonares solitários utilizando Índice de Simpson e Máquina de Vetores de Suporte**. Dissertação (Mestrado). Universidade Federal do Maranhão. São Luís, p. 140. 2009.

SILVA, E. C. et al. Diagnosis of solitary lung nodules using the local form of Ripley's K function. **Computer methods and Programs in Biomedicine**, v. 90, p. 230-239, 2007.

SIMPSON, E. H. Measurement to diversity. **Nature**, 1949.

SONKA, M.; HLAVAC, V.; BOYLE, R. Image processing, analysis and machine vision, 1998. Disponível em: <<http://www.icaen.uiowa.edu/~dip/LECTURE/lecture.html>>. Acesso em: 05 jan. 2012.

SOUSA, J. R. F. S. **Metodologia para detecção automática de nódulos pulmonares**. Dissertação (Mestrado). Universidade Federal do Maranhão. São Luís, p. 98. 2007.

SPSS , 2010. Disponível em: <<http://www.spss.com.br/>>.

VAPNIK, V. N. **Statistical learning theory**. Nova York: Wiley-Interscience, 1998.

VARUTBANGKUL, E. et al. **Combining Boundaries and Ratings from Multiple Observers for Predicting Lung Nodule Characteristics**. International Conference on Biocomputation, Bioinformatics, and Biomedical Technologies. Bucharest:[s.n.]. 2008. p. 82-87.