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

PÉTERSON MORAES DE SOUSA CARVALHO

CLASSIFICAÇÃO DE TECIDOS DA MAMA A PARTIR DE IMAGENS MAMOGRÁFICAS EM MASSA E NÃO MASSA USANDO ÍNDICE DE DIVERSIDADE DE MCINTOSH E MÁQUINA DE VETORES DE SUPORTE

São Luís
2012

PÉTERSON MORAES DE SOUSA CARVALHO

CLASSIFICAÇÃO DE TECIDOS DA MAMA A PARTIR DE IMAGENS MAMOGRÁFICAS EM MASSA E NÃO MASSA USANDO ÍNDICE DE DIVERSIDADE DE MCINTOSH E MÁQUINA DE VETORES DE SUPORTE

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão como parte dos requisitos necessários para obtenção do título de Mestre em Engenharia de Eletricidade na área de concentração Ciência da Computação.

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

São Luís
2012

Carvalho, Péterson Moraes de Sousa.

Classificação de tecidos da mama a partir de imagens mamográficas em massa e não massa usando Índice de Diversidade de McIntosh e máquinas de vetores de suporte / Péterson Moraes de Sousa Carvalho – São Luís, 2012.

75 f.

Orientador: Anselmo Cardoso de Paiva.

Co-orientador: Aristófanes Corrêa Silva.

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

1. Mamografia – Índice de Diversidade de McIntosh. 2. Máquina de vetores de suporte. 3. Reconhecimento de padrões. I. Título.

CDU 621.386.84:618.19-073

CLASSIFICAÇÃO DE TECIDOS DA MAMA A PARTIR DE IMAGENS MAMOGRÁFICAS EM MASSA E NÃO MASSA USANDO ÍNDICE DE DIVERSIDADE DE MCINTOSH E MÁQUINA DE VETORES DE SUPORTE

Péterson Moraes de Sousa Carvalho

Dissertação aprovada em 20 de abril de 2012.

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

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

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

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

Aos meus pais, irmãos e sobrinhos.

AGRADECIMENTOS

À Deus, por ter me proporcionado vida, saúde, paciência e persistência.

Aos meus orientadores, Dr. Anselmo Cardoso de Paiva e Dr. Aristófanes Corrêa Silva, pela confiança, apoio, paciência, competência e dedicação.

Aos meus colegas de mestrado, pelo apoio, motivação, troca de conhecimentos e pelas ajudas quando necessitei.

À FAPEMA pelo suporte financeiro durante o período do mestrado.

Ao corpo docente da Universidade Federal do Maranhão, pelo conhecimento adquirido ao longo do curso, que me proporcionou a elaboração deste trabalho.

À minha família que é muito importante na minha vida.

À todos que, direta ou indiretamente, contribuíram para a elaboração deste trabalho.

“Obstáculo é aquilo que você enxerga,
quando tira os olhos do seu objetivo.”

Henry Ford

RESUMO

O câncer de mama é o segundo tipo de câncer mais frequente no mundo e o que mais acomete as mulheres. Nos últimos anos, vários Sistemas de Detecção e Diagnóstico auxiliados por Computador (*Computer Aided Detection/Diagnosis*) têm sido desenvolvidos no intuito de auxiliar especialistas da área da saúde na detecção e diagnóstico de câncer, servindo como uma segunda opnião. O objetivo deste trabalho é apresentar uma metodologia de discriminação e classificação de regiões extraídas de mamografias em massa e não massa. Neste estudo, o *Digital Database for Screening Mammography* (DDSM) é usado. Para descrever a textura da região de interesse é aplicado o Índice de Diversidade de McIntosh, comumente usado em ecologia. O cálculo deste índice é proposto em quatro abordagens: através do Histograma, da Matriz de Co-ocorrência de Níveis de Cinza, da Matriz de Comprimentos de Corrida de Cinza e da Matriz de Comprimentos de Lacuna de Cinza. Para classificação das regiões em massa e não massa, é utilizado o classificador supervisionado *Support Vector Machine* (SVM). A metodologia apresenta resultados promissores para a classificação de massas e não massas, alcançando uma acurácia de 93,68%.

Palavras-chave: Mamografia, Índice de Diversidade de McIntosh, Máquina de Vetores de Suporte, Reconhecimento de Padrões.

ABSTRACT

Breast cancer is the second most common in the world and which more affects women. In recent years, several Computer Aided Detection/Diagnosis Systems has been developed in order to assist health specialists in the detection and diagnosis of cancer, serving as a second opinion. The aim of this paper is to present a methodology for discrimination and classification of regions extracted from mammograms in mass and non-mass. In this study, Digital Database for Screening Mammography (DDSM) is used. To describe the texture of the region of interest is applied McIntosh Diversity Index, commonly used in ecology. The calculation of this index is proposed in four approaches: through the Histogram, through the Gray Level Co-occurrence Matrix, through the Gray Level Run Length Matrix and through the Gray Level Gap Length Matrix. For the classification of regions in mass and non-mass, is used the supervised classificator Support Vector Machine (SVM). The methodology shows promising results for the classification of masses and non-masses, reaching an accuracy of 93,68%.

Keywords: Mammography, McIntosh Diversity Index, Support Vector Machine, Pattern Recognition.

Artigos Científicos Aceitos para Publicação

CARVALHO, P. M. de S., PAIVA, A. C., SILVA, A. C. Classification of Breast Tissues in Mammographic Images in Mass and Non-Mass using McIntosh's Diversity Index and SVM. International Conference on Machine Learning and Data Mining (MLDM), Berlin, 2012.

SUMÁRIO

1 INTRODUÇÃO 15
1.1 Trabalhos Relacionados 17
1.2 Organização do Trabalho 19
2 FUNDAMENTAÇÃO TEÓRICA 21
2.1 Câncer de Mama 21
2.1.1 Mamografia 22
2.2 Índice de Diversidade 24
2.2.1 Índice de Diversidade de McIntosh 25
2.3 Processamento de Imagens Digitais 26
2.3.1 Quantização Uniforme 28
2.3.2 Análise de Textura 30
2.3.2.1. Histograma 31
2.3.2.2. Matriz de Co-Ocorrência de Níveis de Cinza - GLCM 32
2.3.2.3. Matriz de Comprimentos de Corrida de Cinza - GLRLM 33
2.3.2.4. Matriz de Comprimentos de Lacuna de Cinza - GLGLM 35
2.3.3 Equalização de histograma 36
2.4 Reconhecimento de Padrões 38
2.4.1 Máquina de Vetores de Suporte 40
2.5 Métricas de Validação de Resultados 44
3 METODOLOGIA 45
3.1 Metodologia Proposta 45
3.1.1 Aquisição das Amostras 46
3.1.2 Pré-Processamento 47
3.1.3 Extração de Características 48
3.1.3.1. Índice de Diversidade de McIntosh a partir do Histograma 48
3.1.3.2. Índice de Diversidade de McIntosh a partir da GLCM 49
3.1.3.3. Índice de Diversidade de McIntosh a partir da GLRLM 50
3.1.3.4. Índice de Diversidade de McIntosh a partir da GLGLM 50
3.1.4 Reconhecimento de Padrões 52
3.1.5 Validação de Resultados 54
4 RESULTADOS E DISCUSSÃO 55
4.1 Resultados Obtidos 55
4.1.1 Abordagem usando Histograma 56
4.1.2 Abordagem usando GLCM 57
4.1.3 Abordagem usando GLRLM 59
4.1.4 Abordagem usando GLGLM 61
4.1.5 Abordagem usando GLRLM e GLCM 62
4.1.6 Abordagem usando GLRLM e GLGLM 64

4.2 Resultados Finais 65
4.2.1 Comparação com outros trabalhos relacionados 67
5 CONCLUSÃO 68
REFERÊNCIAS 70
APÊNDICE A – Comparando o uso de amostras de tamanhos diferentes com o uso de amostras de tamanhos iguais 75

LISTA DE FIGURAS

Figura 2.1: (a) Mamografia com Incidência Médio-Lateral (ambas as mamas); (b) Mamografia com Incidência Crânio-Caudal (ambas as mamas). Fonte: (MAMOWEB, 2011)..... 23

Figura 2.2: Mamógrafos. (a) Esquema. Fonte (ACS - American Cancer Society, 2012). (b) Mamógrafo real. Fonte: (MAMOWEB, 2011)..... 23

Figura 2.3: Representação de uma comunidade de três espécies, de acordo com McIntosh. O ponto P representa a comunidade, os eixos representam as espécies. 25

Figura 2.4: Etapas fundamentais em processamento de imagens digitais. Adaptado de (GONZALEZ e WOODS, 2000)..... 27

Figura 2.5: Ilustração da quantização. (a) 16 níveis de cinza; (b) 4 níveis de cinza; e (c) 2 níveis de cinza. Fonte: (JAHNE, 2005)..... 29

Figura 2.6: Exemplos de texturas. 30

Figura 2.7: Exemplo de uma co-ocorrência dos níveis de cinza i e j, com vizinhança $d = 4$, alinhados na horizontal $(\theta = 0)$..... 32

Figura 2.8: (a) Imagem de $M\ x\ N$ pixels. (b) Matriz de Co-ocorrência de Níveis de Cinza da imagem $(d = 2, \theta = 0o)$..... 33

Figura 2.9: Exemplo de uma corrida de nível de cinza i, de comprimento 10 e direção horizontal. 33

Figura 2.10: (a) Imagem de $M\ x\ N$ pixels. (b) Matriz de Comprimentos de Corrida de Cinza da imagem $(\theta = 0o)$. 34

Figura 2.11: Lacuna de nível de cinza g, de comprimento l e direção horizontal. 35

Figura 2.12: (a) Imagem de $M\ x\ N$ pixels. (b) Matriz de Comprimentos de Lacuna de Cinza da imagem $(\theta = 0o)$. 36

Figura 2.13: Equalização do histograma. (a) Imagem original, (b) Histograma da imagem original, (c) Imagem equalizada e (d) Histograma da imagem equalizada. Adaptado de (MATARREDONA, 1994) 37

Figura 2.14: Classificação supervisionada. Fonte: (LORENA e CARVALHO, 2007) 39

Figura 2.15: Superfícies de decisão. (a) linerares e (b) não-linerares. Fonte: (SANTOS, 2002) 39

Figura 2.16: (a) Um hiperplano de separação com margem pequena. (b) Um hiperplano de margem máxima. Fonte: (SANTOS, 2002)..... 40
Figura 2.17: Hiperplano de separação para o caso linearmente separável. Os vetores de suporte estão circulados. Fonte: (SANTOS, 2002) 41
Figura 2.18: Hiperplano de separação para o caso não linearmente separável. Fonte: (SANTOS, 2002)..... 42
Figura 2.19: Mapeamento do espaço de entrada para o espaço de características via função kernel. 43
Figura 3.1: Etapas da metodologia proposta. 45
Figura 3.2: Regiões extraídas de mamografias da base DDSM. (a) massas, (b) não massas... 47
Figura 3.3: (a) Massa original, (b) Massa após a equalização do histograma. 48
Figura 3.4: Definição da entidade espécie (e) na imagem, para cada abordagem. (a) Entidade espécie é o nível de cinza, (b) Entidade espécie é o par de pixel (transição de i para j), (c) Entidade espécie é a corrida de cinza de comprimento j e (d) Entidade espécie é a lacuna de cinza de tamanho l. 52

LISTA DE TABELAS

Tabela 4.1: Parâmetros usados pelo classificador SVM, obtidos a partir das características extraídas com o Índice de Diversidade de McIntosh calculado a partir do histograma ... 56
Tabela 4.2: Resultados obtidos na classificação de amostras em massa e não massa utilizando o Índice de Diversidade de McIntosh a partir do Histograma. ... 57
Tabela 4.3: Parâmetros usados pelo classificador SVM, obtidos a partir das características extraídas com o Índice de Diversidade de McIntosh calculado a partir da GLCM ... 58
Tabela 4.4: Resultados obtidos na classificação de amostras em massa e não massa utilizando o Índice de Diversidade de McIntosh a partir da GLCM. ... 58
Tabela 4.5: Parâmetros usados pelo classificador SVM, obtidos a partir das características extraídas com o Índice de Diversidade de McIntosh calculado a partir da GLRLM ... 60
Tabela 4.6: Resultados obtidos na classificação de amostras em massa e não massa utilizando o Índice de Diversidade de McIntosh a partir da GLRLM. ... 60
Tabela 4.7: Parâmetros usados pelo classificador SVM, obtidos a partir das características extraídas com o Índice de Diversidade de McIntosh calculado a partir da GLGLM ... 61
Tabela 4.8: Resultados obtidos na classificação de amostras em massa e não massa utilizando o Índice de Diversidade de McIntosh a partir da GLGLM. ... 62
Tabela 4.9: Parâmetros usados pelo classificador SVM, obtidos a partir das características extraídas com o Índice de Diversidade de McIntosh calculado a partir da GLRLM e GLCM ... 63
Tabela 4.10: Resultados obtidos na classificação de amostras não equalizadas, utilizando o Índice de Diversidade de McIntosh a partir da GLRLM e GLCM. ... 63
Tabela 4.11: Parâmetros usados pelo classificador SVM, obtidos a partir das características extraídas com o Índice de Diversidade de McIntosh calculado a partir da GLRLM e GLGLM ... 64
Tabela 4.12: Resultados obtidos na classificação de amostras não equalizadas, utilizando o Índice de Diversidade de McIntosh a partir da GLRLM e GLGLM. ... 65
Tabela 4.13: Acurácia máxima obtida em cada abordagem empregada neste trabalho. ... 65
Tabela 4.14: Comparação com alguns trabalhos referentes à classificação de tecidos extraídos de mamografias em massa e não massa. ... 67
Tabela A.1: Comparação dos melhores resultados obtidos a partir de amostras de tamanhos diferentes e tamanhos padronizados. ... 75

LISTA DE ABREVIATURAS E SIGLAS

ACS	*American Cancer Society*
CAD	*Computer-Aided Detection*
CADx	*Computer-Aided Diagnosis*
DDSM	*Digital Database for Screening Mammography*
FN	Falso Negativo
FP	Falso Positivo
GLCM	*Gray Level Co-occurrence Matrix*
GLGLM	*Gray Level Gap Length Matrix*
GLRLM	*Gray Level Run Length Matrix*
INCA	Instituto Nacional do Câncer
MIAS	*Mammograms Image Analysis Society*
OMS	Organização Mundial da Saúde
RBF	*Radial Basis Function*
SVM	*Support Vector Machine*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

1 INTRODUÇÃO

O câncer representa um grande problema de saúde pública mundial. Corresponde 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 (INCA - Instituto Nacional do Câncer, 2011b). Cânceres não tratados podem causar doenças graves e até morte (ACS - American Cancer Society, 2012). A Organização Mundial da Saúde (OMS) estimou que, no ano 2030, podem-se esperar 27 milhões de casos incidentes de câncer, 17 milhões de mortes por câncer e 75 milhões de pessoas vivas, anualmente, com câncer. O maior efeito desse aumento vai incidir em países de baixa e média renda (INCA - Instituto Nacional do Câncer, 2012).

No Brasil, as estimativas para o ano de 2012, válidas também para o ano de 2013, apontam a ocorrência de aproximadamente 518.510 casos novos de câncer, incluindo os casos de pele não melanoma, reforçando a magnitude do problema do câncer no país. Sem os casos de câncer de pele não melanoma, estima-se um total 385 mil casos novos. Os tipos mais incidentes serão os cânceres de pele não melanoma, próstata, pulmão, cólon e reto e estômago para o sexo masculino; e os cânceres de pele não melanoma, mama, colo do útero, cólon e reto e glândula tireoide para o sexo feminimo (INCA - Instituto Nacional do Câncer, 2012).

O câncer de mama é o segundo tipo mais frequente no mundo, sendo mais comum entre as mulheres, e o que mais as acomete, respondendo por 22% dos casos novos de câncer a cada ano. Se diagnosticado e tratado oportunamente, o prognóstico é relativamente bom (INCA - Instituto Nacional do Câncer, 2011d). Cerca de 1,4 milhões de casos novos dessa neoplasia foram esperados para o ano de 2008 em todo o mundo, o que representa 23% de todos os tipos de câncer. Em 2012, esperam-se, para o Brasil, 52.680 casos novos de câncer de mama, com um risco estimado de 52 casos a cada 100 mil mulheres (INCA - Instituto Nacional do Câncer, 2012).

As regiões brasileiras onde o câncer de mama é o mais incidente entre as mulheres são Sudeste (69/100 mil), Sul (65/100 mil), Centro-Oeste (48/100 mil) e Nordeste (32/100 mil). Na região Norte (19/100 mil) é o segundo tumor mais incidente. No Maranhão, a taxa bruta de incidência estimada de câncer de mama para 2012 indica 13,97 casos para cada 100 mil

mulheres, e na capital, São Luís, esta taxa fica em 35,65 casos para cada 100 mil mulheres (INCA - Instituto Nacional do Câncer, 2012).

As taxas de mortalidade por câncer de mama continuam elevadas no Brasil, muito provavelmente porque a doença ainda é diagnosticada em estágios avançados. A sobrevida após cinco anos na população de países desenvolvidos tem apresentado um discreto aumento, cerca de 85%. Entretanto, nos países em desenvolvimento, a sobrevida fica em torno de 60% (INCA - Instituto Nacional do Câncer, 2012).

É conhecido que a melhor forma de prevenção de câncer de mama é o seu diagnóstico precoce, fato que diminui a mortalidade, aumentando a eficácia do tratamento. A mamografia é, atualmente, a melhor técnica de detecção precoce de lesões não apalpáveis na mama, com altas chances de ser um cancer curável (ACS - American Cancer Society, 2012). Consiste em um exame de raio-x da mama, onde o resultado é a produção, em uma folha de filme, de uma imagem em tons de cinza, que é lida e interpretada por um radiologista.

O diagnóstico baseado em mamografias é uma etapa sensível, uma vez que radiologistas podem fornecer interpretações diferentes para um mesmo exame. Além disso, a interpretação é uma tarefa repetitiva, requerendo um grande nível de atenção sobre os detalhes presentes na imagem. Em decorrência desses motivos, surgiram nos últimos anos um considerável interesse no uso de técnicas de processamento de imagens digitais para melhorar a análise de exames da mama.

O uso conjunto de técnicas de processamento de imagens digitais e reconhecimento de padrões, tornaram viáveis a produção de Sistemas de detecção (CAD – *Computer-Aided Detection*) e diagnóstico (CADx – *Computer-Aided Diagnosis*), com o objetivo de auxiliar o radiologista, indicando áreas suspeitas na mamografia, assim como anormalidades mascaradas, aumentando, desta forma, a precisão do diagnóstico e servindo como uma segunda opnião para o especialisa (JUNIOR, 2008).

Neste trabalho, é apresentado uma metodologia de Diagnóstico Auxiliado por Computador (CADx) para a classificação de regiões em imagens mamográficas em duas classes: massa e não massa. São consideradas massas todas as regiões da mamografia que correspondem a uma neoplasia maligna ou benigna, e não massas todas as regiões que não são neoplasias. De acordo com (ZHANG e KUMAR, 2006), o objetivo a ser cumprido pelos sistemas ou metodologias CADx é altamente dependente do método de extração de características. Propomos, neste trabalho, a utilização do índice de diversidade de McIntosh como método de extração de características de textura de regiões extraídas de mamografias.

Este índice é utilizado em conjunto com técnicas de processamento de imagens digitais e reconhecimento de padrões (Máquina de Vetores de Suporte) para a classificação de regiões pré-segmentadas da imagem mamográfica em massa e não massa.

O objetivo deste trabalho é avaliar a eficácia do índice de diversidade de McIntosh como método de extração de características de textura de regiões de interesse de imagens mamográficas, de modo que estas regiões possam ser discriminadas em massa e não-massa. Desta forma, pretende-se apresentar uma inovação na extração de textura de imagens médicas, contribuindo, assim, para a literatura da área de sistemas e metodologias CAD/CADx.

1.1 Trabalhos Relacionados

Vários trabalhos têm sido desenvolvidos, fornecendo metodologias eficientes para ajudar na detecção e diagnóstico de câncer de mama.

Em (NUNES, SILVA e PAIVA, 2010) é proposta uma metodologia para detecção de massas em imagens mamográficas através do algorítmo de agrupamento *K-means* e a técnica *Template Matching* para identificação de regiões suspeitas. Em seguida, como atributo de textura da região suspeita, é utilizado o índice de diversidade de Simpson. A informação de textura é, então, usada pela Máquina de Vetores de Suporte (SVM) para classificar as regiões suspeitas em duas classes: massa e não-massa. Este trabalho alcançou uma acurácia de 83,94%, sensibilidade de 83,24% e especificidade de 84,14%.

Em (MOHANTY, BEBERTA e LENKA, 2011) é proposto um sistema para classificação de regiões de interesse de imagens mamográficas em benigno e maligno. O sistema consiste de três etapas. A primeira corresponde à extração das regiões de interesse de 256 x 256 pixels. Na segunda etapa, de extração de características, é utilizado um conjunto de 19 características, calculadas a partir das Matrizes de Co-ocorrência de Níveis de Cinza e Matrizes de Comprimentos de Corrida de Cinza. Na terceira etapa, é empregado a técnica de Mineração de Regras de Associação para classificar as regiões de interesse em benigno e maligno, atingindo uma acurácia de 94,9%. Em uma análise adicional, onde são selecionadas 12 das 19 características, é atingido uma acurácia de 92,3%.

Em (MOAYEDI, AZIMIFAR, *et al.*, 2010) é proposta uma metodologia para classificação automática de tecidos a partir de mamografias. Inicialmente, é realizado um pré-

processamento para extração do músculo peitoral e segmentação da região de interesse. Em seguida, é empregada a Transformada de Contourlet como descritor de características e Algorítmo Genético para selecionar características da base. A classificação é realizada através de *Successive Enhancement Learning* (SEL), *Weighted Support Vector Machine* (WSVM), *Support Vector Machine* (SVM). O trabalho apresenta resultados com acurácia de 96,6%, 91,5% e 82,1% respectivamente para cada uma das técnicas de classificação utilizada, com as imagens da base *Mammograms Image Analysis Society* (MIAS).

Em (SOUSA, 2011), é proposto o uso do índice de diversidade de Shannon como métrica de textura para classificação de regiões de interesse da mama em massa e não massa. O índice é calculado a partir de quatro abordagens independentes: global, onde são considerados todos os pixels da região de interesse, circular, onde são considerados apenas os pixels dentro do círculo concentrico à região, anelar, onde são considerados apenas os pixels dentro do anél concentrico à região, e direcional, onde são considerados pares de pixels de mesma intensidade (correlacionados), de toda a região de interesse, separados por uma certa distância e inclinados sob um certa direção. Foram consideradas as direções 0^o, 45^o, 90^o e 135^o, e uma tolerância angular de $22,5^o$. Para classificar os tecidos da mama em massa e não-massa, foi adotado a Máquina de Vetores de Suporte. A melhor abordagem constatada foi a direcional, na qual foi alcançado uma acurácia de 99,88%, uma sensibilidade de 99,94% e uma especificidade de 99,78%.

Em (MERT, KILIC e AKAN, 2011) é proposta a classificação de massas da mama em benigno e maligno. A base de imagens utilizada foi a *Wisconsin Diagnostic Breast Cancer* (WDBC), onde cada instancia consiste de 30 características. Das 30 características, apenas 2 foram selecionadas, através do algorítimo de redução de dimensionalidade *Independent Component Analysis* (ICA). A Máquina de Vetores de Suporte foi utilizada para classificar as amostras, com uso das funções de núcleo polinomial e de base radial (RBF), onde os resultados da classificação são apresentados através de curvas *Receiver Operating Characteristics* (ROC). Foi atingido uma acurácia máxima de 94,41%.

Em (GULIATO, DE OLIVEIRA e TRAINA, 2010) é proposta uma abordagem baseada nas curvas de Hilbert para classificar massas da mama como benignas ou malignas. Para a extração de características de textura, foi empregado a descrição da forma do contorno da massa através de curvas de Hilbert e *quad-regions*, onde a imagem é dividida em sub-regiões, que são processadas de forma independentes. Foi usado um conjunto de 111 contornos de 65 massas benignas e 46 massas malignas. Uma Rede Neural Artificial (NN) foi utilizada para

classificar tumores da mama em benigno e maligno, sendo atingida uma acurácia de 99%. Os resultados da classificação foram avaliados usando o valor da área sobre a curva *Receiver Operating Characteristic* (ROC).

Em (MARTINS, JUNIOR, *et al.*, 2010) é comparado o desempenho dos classificadores Máquina de Vetores de Suporte (SVM) e Rede Neural Bayesiana na tarefa de classificação de regiões extraídas de mamografias em tecidos normais (não massas) e anormais (massas), baseados em características calculadas através de semivariograma, levando em consideração a distribuição espacial de informações introduzida, inicialmente, neste contexto em (SILVA, CARVALHO e GATTASS, 2005). As variáveis são inicialmente selecionadas usando *Stepwise*. O melhor resultado de classificação alcançado pelo SVM aponta 86,11% de acurácia, contra 76,85% de acurácia atingida com as Redes Neurais Bayesianas.

Podemos verificar, nos trabalhos citados, que metodologias baseadas em características de textura e reconhecimento de padrões apresentam resultados promissores na detecção de câncer de mama, baseado em mamografias. Verifica-se, também, que a classificação de regiões de interesse de mamografias, em massa e não massa, representa uma etapa crucial nas metodologias de detecção de câncer de mama.

1.2 Organização do Trabalho

O restante deste trabalho está organizado em mais quatro capítulos, descritos resumidamente a seguir.

No Capítulo 2, é apresentada a fundamentação teórica necessária para a compreensão deste trabalho. São descritos os conceitos de câncer de mama, mamografia, índice de diversidade de McIntosh, processamento de imagens digitais, quantização uniforme, equalização de histograma, análise de textura (histograma, GLCM, GLRLM e GLGLM), reconhecimento de padrões (Máquina de Vetores de Suporte) e métricas de validação dos resultados.

No Capítulo 3 é descrita a metodologia utilizada para realizar a classificação de regiões de interesse, extraídas de mamografias digitais, em massa e não massa, utilizando a extração de características de textura baseada no conceito de diversidade ecológica (índice de diversidade de McIntosh) e o reconhecimento de padrões através de Máquina de Vetores de Suporte.

No Capítulo 4 são apresentados e discutidos os resultados obtidos através da metodologia proposta. O Capítulo 5 apresenta a conclusão deste trabalho, mostrando a eficiência dos métodos utilizados e apresentando sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo é apresentada a fundamentação teórica necessária para a compreensão da metodologia utilizada neste trabalho. Aborda-se o câncer de mama, a mamografia, Índice de Diversidade de McIntosh, processamento de imagens digitais, quantização uniforme, análise de textura (Histograma, Matriz de Co-ocorrência de Nível de Cinza, Matriz de Comprimentos de Corrida de Cinza, e Matriz de Comprimentos de Lacuna de Cinza), equalização de histograma, método de classificação e reconhecimento de padrão utilizando Máquina de Vetores de Suporte e as métricas de validação de resultados.

2.1 Câncer de Mama

O câncer de mama é um tumor maligno que começa nas células da mama. Um tumor maligno é um grupo de células cancerosas que podem crescer em tecidos circundantes ou se espalhar (metástase) para áreas distantes do corpo. Os homens podem desenvolver câncer de mama, mas a doença ocorre quase que inteiramente nas mulheres (ACS - American Cancer Society, 2012). O sintoma mais comum de câncer de mama é o aparecimento de um nódulo, geralmente indolor, duro e irregular, mas há tumores que são de consistência branda, globulosos e bem definidos (INCA - Instituto Nacional do Câncer, 2011a).

O câncer de mama, assim como outras doenças, apresenta alguns fatores de risco[1]: histórico familiar de câncer de mama (principalmente em parentes de primeiro grau antes dos 50 anos), menarca precoce (idade da primeira menstruação menor que 12 anos), menopausa tardia (após os 50 anos), primeira gravidez após os 30 anos, nuliparidade (ausência de gestação), terapia de reposição hormonal pós-menopausa (principalmente se prolongada por mais de cinco anos), ingestação regular de bebida alcoólica, obesidade e sedentarismo (INCA - Instituto Nacional do Câncer, 2011a). No entanto, ter um fator de risco, ou mesmo vários, não significa que você vai ter a doença. A maioria das mulheres que têm um ou mais fatores de risco de câncer de mama, nunca desenvolveram a doença, enquanto muitas mulheres com câncer de mama não possuem fatores de risco aparente. Mesmo quando uma mulher com

[1] Um fator de risco é qualquer coisa que afete a sua chance de adquirir uma doença, como câncer (ACS - American Cancer Society, 2012).

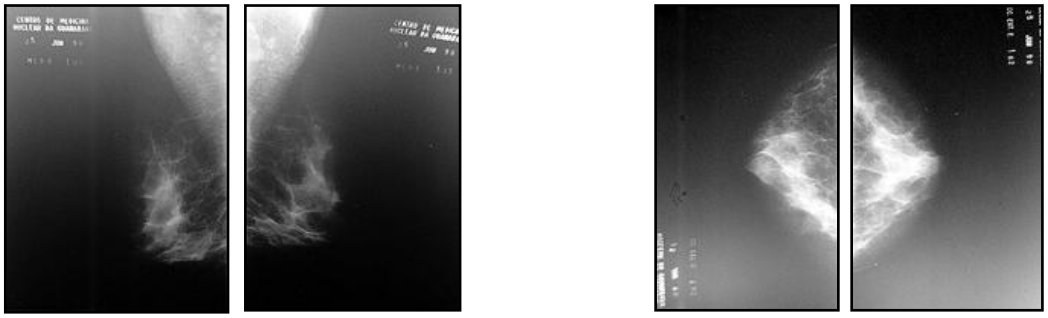
fatores de risco desenvolve o câncer de mama, é difícil saber o quanto esses fatores podem ter contribuído para a doença (ACS - American Cancer Society, 2012).

Para o tratamento ser eficiente (ter maiores chances de ser bem sucedido) é necessário que o câncer seja detectado ainda no início. As formas mais eficazes para a detecção precoce do câncer de mama são o exame clínico e a mamografia. O exame clínico das mamas (ECM), realizado por um médico ou enfermeira treinados, pode detectar tumor de até 1 (um) centímetro, se superficial. Deve ser feito uma vez por ano pelas mulheres entre 40 e 49 anos. A mamografia permite a identificação de lesões em fase inicial, muito pequenas (medindo milímetros). Deve ser realizada a cada dois anos por mulheres entre 50 e 69 anos, ou segundo recomendação médica (INCA - Instituto Nacional do Câncer, 2011c).

2.1.1 Mamografia

A mamografia é um exame utilizado para detectar e avaliar anormalidades na mama, tanto em mulheres que não têm queixas ou sintomas quanto em mulheres que têm sintomas de doenças mamárias (ACS - American Cancer Society, 2012). É, atualmente, a melhor técnica de detecção prévia de lesões não apalpáveis na mama com altas chances de ser um câncer curável. Há, basicamente, dois tipos de exames: a mamografia de rotina (rastreamento) e a mamografia de diagnóstico. A mamografia de rotina é a mais frequentemente usada, com a finalidade de procurar por câncer em mulheres que não apresentam nenhum sintoma. Ela é proposta apenas para mulheres acima dos quarenta anos de idade (AZEVEDO e PEIXOTO, 1993). A mamografia de diagnóstico é utilizada para diagnosticar a doença de mama em mulheres que apresentam sintomas, como nódulo ou secreção do mamilo, ou um resultado anormal em uma mamografia. Ela inclui imagens da área de preocupação (ACS - American Cancer Society, 2012).

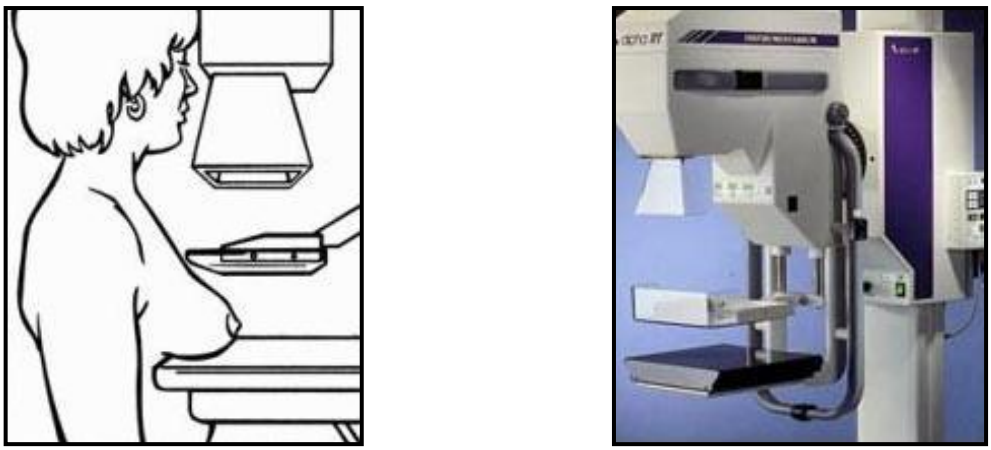
A mamografia consiste em uma radiografia da mama, tomada, geralmente, em duas visões (imagens de Raio-X tiradas de ângulos diferentes) de cada mama: Médio-Lateral Oblíqua (Figura 2.1a) e Crânio-Caudal (Figura 2.1b). Este procedimento produz uma imagem em tons de cinza do tecido mamário ou em uma folha grande de filme ou como uma imagem digital de computador, que é lida, ou interpretada, por um radiologista (ACS - American Cancer Society, 2012).



(a) (b)

Figura 2.1: (a) Mamografia com Incidência Médio-Lateral (ambas as mamas); (b) Mamografia com Incidência Crânio-Caudal (ambas as mamas). Fonte: (MAMOWEB, 2011)

O exame de mamografia é realizado por uma máquina especial chamada mamógrafo (Figura 2.2), onde a mama é comprimida entre duas placas plásticas (Figura 2.2a), anexas ao aparelho, de modo que os tecidos da mama fiquem dispersos e com uma espessura uniforme, fornecendo, desta forma, uma imagem mais nítida. Orientações rigorosas garantem que o mamógrafo é seguro e utiliza a menor dose de radiação possível. A compressão, embora cause desconforto, dura alguns segundos, e todo o processo leva 20 minutos (ACS - American Cancer Society, 2012).



(a) (b)

Figura 2.2: Mamógrafos. (a) Esquema. Fonte (ACS - American Cancer Society, 2012). (b) Mamógrafo real. Fonte: (MAMOWEB, 2011)

Embora represente a melhor forma de detecção precoce de câncer de mama, a mamografia também têm algumas limitações. Pode perder alguns tipos de câncer, e isso, às

vezes, leva o acompanhamento de resultados que não são câncer, incluíndo biópsias[2]. Outra limitação é que a mamografia não funciona tão bem em mulheres mais jovens, geralmente porque suas mamas são densas e podem ocultar um tumor. Uma vez que a maioria dos cânceres de mama ocorrem em mulheres mais velhas, isso geralmente não é uma grande preocupação (ACS - American Cancer Society, 2012).

Considerando que alguns fatores, como a sobreposição de tecidos, variedade de formas das lesões e a alta taxa de ruído presentes em imagens mamográficas, dificultam a análise de mamografias (MASCARO, 2007), o que pode gerar uma quantidade relativamente grande de diagnósticos falso-positivos, e que radiologistas podem fornecer diagnósticos diferentes para um mesmo exame (JUNIOR, 2008), ferramentas de apoio aos radiologistas através do computador, CAD/CADx (Computer Aided Detection/Computer Aided Diagnosis), têm sido desenvolvidas nos últimos anos para melhorar o desempenho da análise de mamografias, através da identificação de lesões e classificação de regiões ou de objetos de interesse. O uso dessas ferramentas se tornou uma prática clínica bem aceita para auxiliar radiologistas na interpretação de mamografias (MELLO-THOMS, 2007), servindo como uma segunda opnião. A abordagem mais comum para o desenvolvimento de sistemas CAD/CADx envolve procedimentos de extração de características, realizada tanto por um sistema de computador ou manualmente pelos radiologistas (PAPADOPOULOS, FOTIADIS e LIKAS, 2005).

2.2 Índice de Diversidade

O estudo da diversidade é usado em ecologia para informar a variedade de espécies presentes em uma comunidade ou área. Uma comunidade é definida como um conjunto de espécies que ocorrem em um determinado lugar e tempo (MAGURRAN, 2004). O uso de índices, embora não representem a composição total de uma comunidade, permite dimensionar a riqueza, a igualdade e a diversidade das espécies nos diferentes ambientes estudados, sendo úteis para monitorar e predizer mudanças ambientais. Foram desenvolvidos, inicialmente, para estudos de macroecologia (KENNEDY e SMITH, 1995).

O conceito de diversidade envolve dois parâmetros: **riqueza**, que corresponde à quantidade de espécies, e **abundância relativa**, que é a quantidade de indivíduos de

[2] A mamografia não pode provar que uma área anormal é câncer, de modo que, para confirmar a presença do câncer, uma pequena quantidade de tecido deve ser removido e analisado sob um microscópio (ACS - American Cancer Society, 2012).

determinada espécie, que ocorre em um local ou amostra (PIANKA, 1994). Comunidades com a mesma riqueza podem diferir em diversidade dependendo da distribuição de indivíduos entre as espécies (MCINTOSH, 1967).

O cálculo do índice de diversidade, qualquer que seja, resulta em um único número. De acordo com Mahafee (MAHAFEE e KLOEPPER, 1997), é vantajoso o fato do índice de diversidade utilizar um único número para representar uma determinada situação, uma vez que facilita a comparação em experimentação, assim como possibilita a elucidação de mudanças que ocorrem nas comunidades relacionadas.

2.2.1 Índice de Diversidade de McIntosh

No índice de diversidade proposto por McIntosh uma comunidade pode ser encarada como um ponto em um hipervolume S-dimensional (MAGURRAN, 2004). Cada espécie é teoricamente representada por um eixo em tal espaço hipotético (MCINTOSH, 1967). Se uma comunidade apresenta, por exemplo, três espécies, então ela aparecerá como um ponto em um espaço tridimensional, onde as coordenadas deste ponto serão as abundâncias relativas das espécies (Figura 2.3).

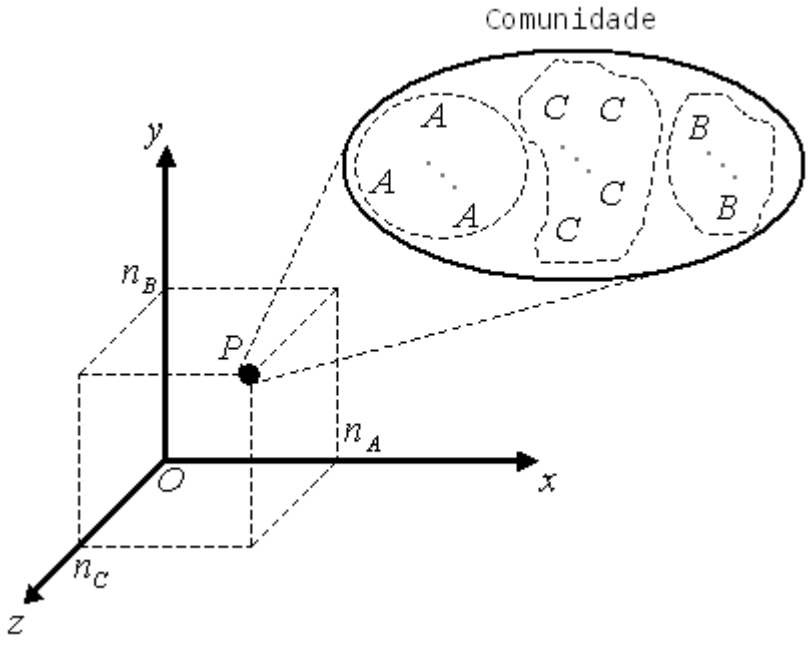


Figura 2.3: Representação de uma comunidade de três espécies, de acordo com McIntosh. O ponto P representa a comunidade, os eixos representam as espécies.

A distância euclidiana da comunidade até a origem pode ser usada como uma medida de diversidade (MAGURRAN, 2004). Essa distância é conhecida como U, sendo calculada como:

$$U = \sqrt{\sum_{i=1}^{s} n_i^{\,2}} \tag{1}$$

onde s é o número de espécies (riqueza) e n_i é o número de indivíduos (abundância relativa) da espécie i. Formalmente, a diversidade de qualquer amostra (MCINTOSH, 1967) é dada por:

$$N - U \tag{2}$$

onde $N = \sum_{i=1}^{s} n_i$ representa o número total de indivíduos da amostra. O valor da diversidade observada aumenta quando o tamanho da amostra (N) aumenta, sendo útil apenas se amostras de mesmo tamanho são comparadas (MCINTOSH, 1967). Um outro índice de diversidade proposto por McIntosh, independente de N, é dado por:

$$\frac{N - U}{N - \sqrt{N}} \tag{3}$$

Este índice tem a vantagem de expressar a diversidade observada como uma proporção da diversidade máxima absoluta, $N - \sqrt{N}$, em um dado N e varia de 0, se houver apenas uma espécie, para 1, se a diversidade é máxima (MCINTOSH, 1967). Este índice é útil quando amostras de tamanhos diferentes são comparadas.

2.3 Processamento de Imagens Digitais

Uma imagem digital é uma imagem $f(x, y)$ discretizada tanto em coordenadas espaciais quanto em brilho. Uma imagem digital pode ser considerada como sendo uma matriz cujos índices de linhas e de colunas identificam um ponto na imagem, e o correspondente valor do elemento da matriz identifica o nível de cinza naquele ponto. Os elementos dessa matriz digital são chamados de elementos da imagem, elementos da figura, *pixels* ou *pels*, estes dois últimos, abreviações de *picture elements* (elementos da figura) (GONZALEZ e WOODS, 2000). Uma imagem em tons de cinza apresenta um único valor de

intensidade associado a cada pixel, enquanto uma imagem colorida possui três valores associados ao pixel: um para o vermelho, um para o verde e um para o azul.

O processamento de imagens digitais compreende processos cujas entradas e saídas são imagens e, além disso, engloba os processos de extração de atributos a partir de imagens, incluindo o reconhecimento de objetos individuais (GONZALEZ e WOODS, 2002). Segundo (GONZALEZ e WOODS, 2000), o interesse em métodos de processamento de imagens digitais objetiva a melhoria de informação visual para a interpretação humana e o processamento de dados de cenas para a percepção automática através de máquinas. Alguns exemplos de aplicação do processamento de imagens digitais são: a análise de recursos naturais e meteorologia por meio de imagens de satélites, análise de imagens biomédicas, aplicações em automação industrial envolvendo o uso de sensores visuais em robôs, etc (JUNIOR, 2008).

O processamento de imagens digitais compreende várias etapas. A Figura 2.4 representa o esquema clássico dessas etapas. Uma vez definido e delimitado o problema, segue-se as etapas de aquisição das imagens digitais, pré-processamento, segmentação, representação e descrição, reconhecimento e interpretação. O conjunto de resultados gerados por uma etapa é utilizada na etapa seguinte. O resultado, ao final do processamento, pode ser ou não representado por uma imagem digital.

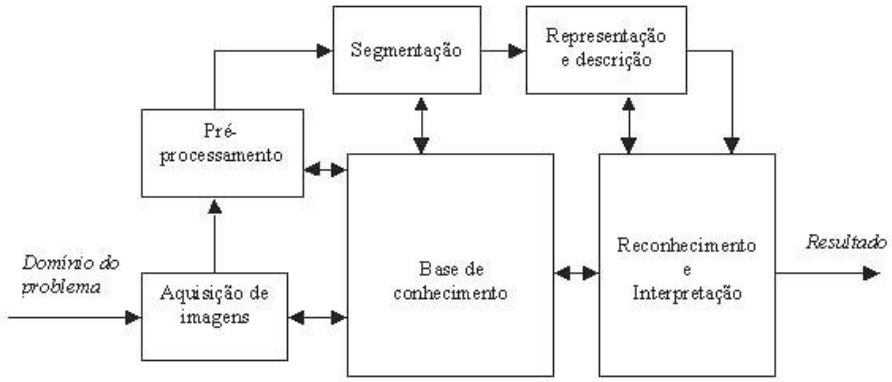


Figura 2.4: Etapas fundamentais em processamento de imagens digitais. Adaptado de (GONZALEZ e WOODS, 2000)

A primeira etapa no processamento de imagens digitais é a **aquisição de imagens**, onde um digitalizador converte a imagem analógica para digital.

A etapa seguinte é o **pré-processamento da imagem**, que visa melhorar a imagem, através de técnicas como realce de contraste, remoção de ruído, filtros morfológicos, dentre outras, de forma a aumentar as chances para o sucesso dos processos seguintes.

A terceira etapa, **segmentação**, procura desmembrar a imagem em seus componentes básicos (objetos), de acordo com suas características. É um processo de particionamento da imagem em regiões desconexas, onde os elementos de uma mesma região devem ser o mais homogêneo possível e os elementos de regiões distintas o mais heterogêneo possível. Não existe na literatura um método geral de segmentação, que se aplica a todas as categorias de imagens. De acordo com a complexidade envolvida no problema, a segmentação pode ser abordada de três formas: manual, semi-automática e automática.

Na segmentação manual o processo de separação de um objeto de interesse é realizado por um especialista humano, com o uso de ferramentas que o auxiliam de forma visual. Na segmentação semi-automática, o especialista passa informações a respeito do que buscar na imagem, ou onde buscar determinada característica, para um algorítmo capaz de processá-las e, assim, poder realizar a segmentação. A mais complexa de todas, por não apresentar conhecimento a priori do que buscar na imagem, é a segmentação automática, que precisa ser robusta para conseguir separar as várias regiões (ou objetos) da imagem em conjuntos desconexos, obdecendo aos critérios de similaridade entre cada região.

A quarta etapa, **representação e descrição**, também conhecida como extração de características, tem por objetivo extrair características que resultem em alguma informação quantitativa de interesse ou que sejam básicas para discriminação entre classes de objetos. O conjunto dessas medidas compõe um vetor de características que define um padrão para uma determinada área de interesse.

A última etapa envolve **reconhecimento e interpretação**. O reconhecimento busca atribuir um rótulo a um objeto, baseado na informação (padrão) fornecida pelo seu descritor (vetor de característica), enquanto a interpretação atribui um significado a um conjunto de objetos reconhecidos.

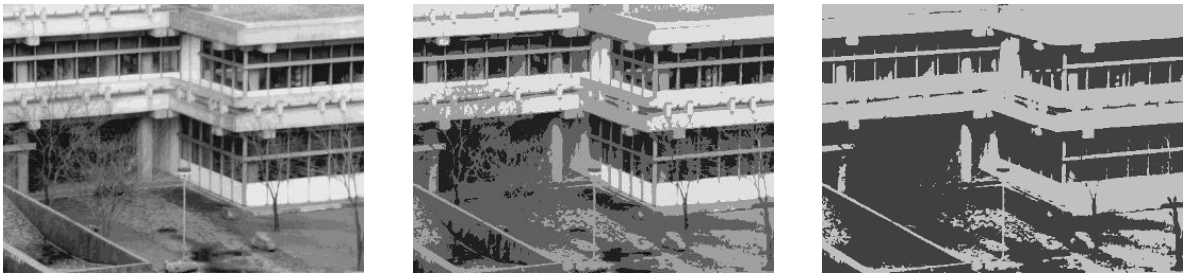
2.3.1 Quantização Uniforme

De acordo com (GONZALEZ e WOODS, 2000), para ser adequada para processamento digital, uma imagem contínua $f(x, y)$ precisa ser digitalizada tanto espacialmente quanto em amplitude (nível de cinza). A digitalização (discretização) das coordenadas espaciais (x, y) é

denominada amostragem da imagem e a digitalização da amplitude é chamada quantização em níveis de cinza. Em (PEDRINI e SCHWARTZ, 2008), a quantização consiste em escolher o número inteiro L de níveis de cinza (em uma imagem monocromática) permitidos para cada ponto da imagem, onde $L = 2^b$, sendo b o número de bits necessário para armazenar a imagem digitalizada. Assim, dada uma imagem com L níveis de cinza, se houver necessidade de quantizá-la para L' níveis de cinza, onde $L' < L$, podemos usar a quantização uniforme, que consiste em dividir a escala de cinza da imagem em intervalos iguais, onde cada intervalo é mapeado para um valor de cinza na imagem quantizada, de modo que a escala de cinza da imagem quantizada é dada por $[0, L' - 1]$. Uma forma de calcular este mapeamento é através da fórmula (SILVA, 2010):

$$q(i,j) = (2^b - 1) \cdot \frac{p(i,j) - I_{min}}{I_{max} - I_{min}} \quad (4)$$

onde $q(i,j)$ é o nível de cinza do pixel (i,j) da nova imagem (quantizada), $p(i,j)$ é o nível de cinza do pixel (i,j) da imagem original, $[I_{min}, I_{max}]$ é a escala de cinza da imagem original, e b é o número de bits necessário para armazenar cada pixel da imagem quantizada. Por exemplo, se uma imagem com níveis de cinza no intervalo $[0, 255]$ for quantizada uniformimente para 4 níveis de cinza (2 bits por pixel), então sua escala de cinza será mapeada para a escala $[0, 3]$. Uma imagem digital degrada à medida que a quantização de níveis de cinza diminui (GONZALEZ e WOODS, 2000). A Figura 2.5 mostra o efeito de perda de informação ocasionado pela quantização em baixos níveis de cinza.

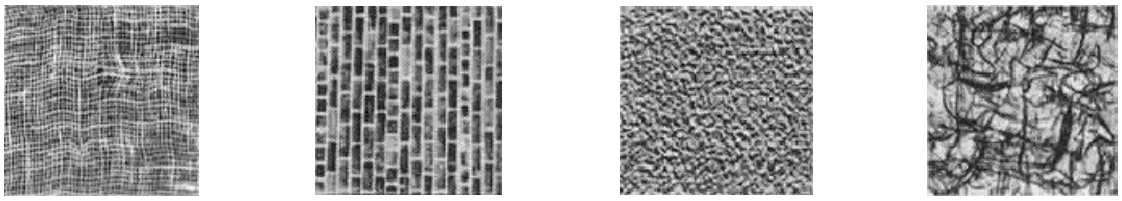


(a) (b) (c)

Figura 2.5: Ilustração da quantização. (a) 16 níveis de cinza; (b) 4 níveis de cinza; e (c) 2 níveis de cinza. Fonte: (JAHNE, 2005)

2.3.2 Análise de Textura

A textura encontra-se entre as características empregadas pelo sistema visual humano, contendo informações sobre a distribuição espacial e a variação de luminosidade, além de descrever o arranjo estrutural das superfícies (Figura 2.6) e relações entre regiões vizinhas (PEDRINI e SCHWARTZ, 2008). Elas ajudam o sistema visual humano a reconhecer objetos, sendo fundamentais no perfeito entendimento de uma cena (MASCARO, 2007).



(a) (b) (c) (d)

Figura 2.6: Exemplos de texturas.

Embora não exista uma definição formal para textura, na literatura são encontradas muitas definições alternativas e métodos para análise de textura. Segundo (HARALICK, 1979), uma textura pode ser descrita pela interação entre as primitivas tonais que a compõem, estas ocorrendo em diferente número e formas. Em (GONZALEZ e WOODS, 2000) a textura é definida como um descritor que intuitivamente fornece medidas de propriedades como suavidade, rugosidade e regularidade. De acordo com (HAWKINS, 1970), a noção de textura parece depender de três ingredientes: (1) alguma "ordem" local é repetida ao longo de uma região que é grande em comparação com o tamanho da ordem, (2) a ordem consiste no arranjo não-aleatório de partes elementares e (3) as partes são entidades mais ou menos uniforme com aproximadamente as mesmas dimensões em todos os lugares dentro da região texturizada. A textura, de um modo geral, pode ser caracterizada como um conceito bidimensional, onde uma dimensão contém as propriedades primitivas da tonalidade e a outra corresponde aos relacionamentos espaciais entre elas (JUNIOR, 2008).

A textura é muitas vezes descrita qualitativamente pela sua aspereza, no sentido de que um pedaço de pano de lã é mais áspero do que um pedaço de seda, sob as mesmas condições de visualização. O índice de aspereza está relacionado com o período de repetição espacial da estrutura local. Um período grande implica uma textura grossa; um período pequeno implica uma textura fina (PRATT, 2001). De modo semelhante, (PEDRINI e SCHWARTZ, 2008) define as texturas finas como sendo a ocorrência de interações aleatórias e grandes variações

no nível de cinza entre as primitivas tonais, e as texturas grossas como interações melhor definidas, com a presença de regiões mais homogêneas.

A análise de textura é uma aplicação relevante em análise de imagens digitais, uma vez que possibilita distinguir regiões da imagem que apresentam as mesmas características de padrões (CONCI, AZEVEDO e LETA, 2008), mas é complicada pelo fato de que ambos os padrões e repetição periódica podem mostrar flutuação aleatória significativa (JAHNE, 2005), como na Figura 2.6c e Figura 2.6d, por exemplo. De acordo com (GONZALEZ e WOODS, 2000), as três abordagens principais usadas em processamento de imagens para a análise de texturas são a estrutural, a espectral e estatística. A abordagem estrutural considera que texturas são compostas de primitivas dispostas de forma aproximadamente regular e repetitiva, conforme regras bem definidas. A abordagem espectral é baseada em propriedades do espectro de Fourier sendo utilizada principalmente na detecção de periodicidade global em uma imagem através da identificação de picos de alta energia no espectro. A abordagem estatística define a textura como um conjunto de medidas locais extraídas do padrão, favorecendo a descrição de imagens através de regras estatísticas que regem tanto a distribuição quanto a relação entre os diferentes níveis de cinza.

Uma vez que, neste trabalho, propomos descrever a textura dos tecidos de regiões de imagens mamográficas através do índice de diversidade de McIntosh, que é uma medida estatística, decidimos explorar a abordagem estatística, através das estatísticas de informação de textura de primeira ordem, como o Histograma, segunda ordem, como a Matriz de Co-ocorrência de Níveis de Cinza (GLCM), e ordem superior, como a Matriz de Comprimentos de Corrida de Cinza (GLRLM) e a Matriz de Comprimentos de Lacuna de Cinza (GLGLM). Para adaptar o conceito de diversidade ecológica, em cada uma dessas estatísticas de informação de textura, foi definido a entidade espécie como sendo a primitiva tonal explorada na imagem.

2.3.2.1. Histograma

Um histograma é uma estatística de primeira ordem que representa a frequência dos níveis de cinza dos pixels na imagem. Em (GONZALEZ e WOODS, 2002), o histograma de uma imagem digital, com níveis de cinza no intervalo $[0, L-1]$, é definida pela função discreta $h(r_k) = n_k$, onde r_k é o k-ésimo nível de cinza e n_k é o número de pixels na imagem com intensidade r_k.

2.3.2.2. Matriz de Co-Ocorrência de Níveis de Cinza - GLCM

Dado um relacionamento espacial entre os pixels componentes de uma textura, os elementos da matriz de co-ocorrência de níveis de cinza (GLCM, do inglês *Gray Level Co-Occurrence Matrix*) descrevem a frequência com que ocorrem as transições de nível de cinza entre pares de pixels (PEDRINI e SCHWARTZ, 2008). Efetuando-se variações na relação espacial, por meio de alterações na orientação e distância entre as coordenadas dos pixels, podem ser obtidas diversas matrizes de co-ocorrência, a partir das quais são extraídas medidas utilizadas para análise de texturas (HARALICK, SHANMUGAM e DINSTEIN, 1973).

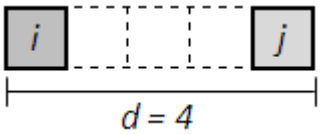


Figura 2.7: Exemplo de uma co-ocorrência dos níveis de cinza i e j, com vizinhança $d = 4$, alinhados na horizontal ($\theta = 0$).

Mais especificamente, dada uma imagem S, com níveis de cinza no intervalo $[0, L-1]$, cada célula (i, j) da matriz de co-ocorrência, com $0 \le i \le L-1$ e $0 \le j \le L-1$, funciona como um contador e armazena a frequência, denotada por $P(i, j, d, \theta)$, com que dois pixels ocorrem na imagem, separados por uma distância d, sob uma direção θ, um com a cor i e outro com a cor j (Figura 2.8). O cálculo do elemento da matriz de co-ocorrência, para as direções 0^o, 45^o, 90^o e 135^o, é descrito através de 4 equações (HARALICK, SHANMUGAM e DINSTEIN, 1973):

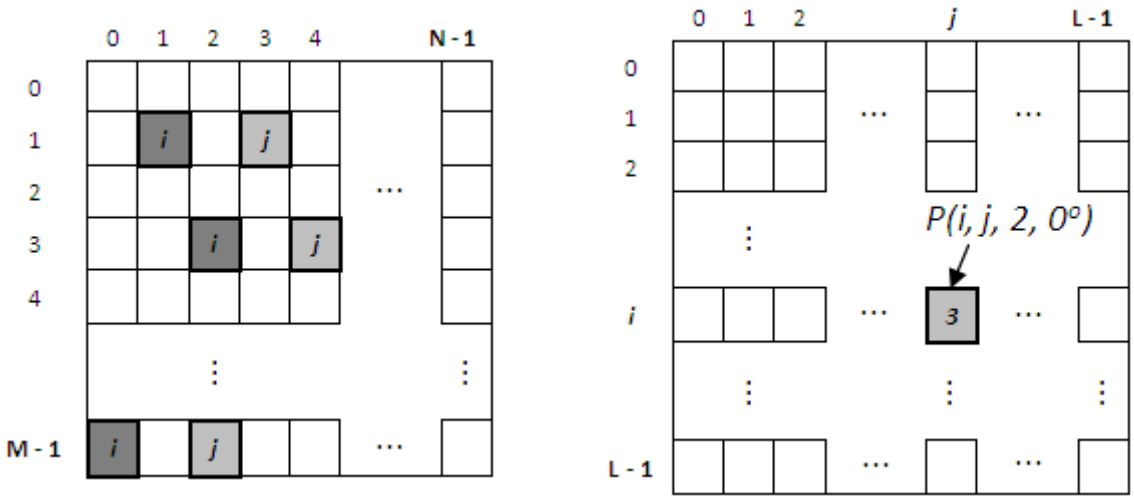
$$P(i, j, d, 0^o) = \#\{((k,l),(m,n)) \mid k-m = 0, |l-n| = d,\ f(k,l) = i,\ f(m,n) = j\} \quad (5)$$

$$P(i, j, d, 45^o) = \#\{((k,l),(m,n)) \mid k-m = d, l-n = -d,\ f(k,l) = i,\ f(m,n) = j\} \quad (6)$$

$$P(i, j, d, 90^o) = \#\{((k,l),(m,n)) \mid |k-m| = d, l-n = 0,\ f(k,l) = i,\ f(m,n) = j\} \quad (7)$$

$$P(i, j, d, 135^o) = \#\{((k,l),(m,n)) \mid k-m = d, l-n = d,\ f(k,l) = i,\ f(m,n) = j\} \quad (8)$$

onde "#" denota o número de pares $((k,l),(m,n))$ do conjunto, e $f(x,y)$ denota a função nível de cinza no pixel (x,y).



(a) (b)

Figura 2.8: (a) Imagem de $M\ x\ N$ pixels. (b) Matriz de Co-ocorrência de Níveis de Cinza da imagem ($d = 2, \theta = 0^o$).

A Figura 2.8b ilustra a estrutura da GLCM, construída a partir da imagem da Figura 2.8a. O tamanho da matriz é $L\ x\ L$, sendo L a quantidade máxima de níveis de cinza que a imagem pode apresentar. Na imagem (Figura 2.8a), por exemplo, há 3 pares de pixels, com vizinhança 2 e alinhamento na horizontal, onde o primeiro pixel tem intensidade i e o segundo tem intensidade j. Assim, a célula (i, j) da GLCM registra a frequência $P(i, j, 2, 0^o) = 3$.

2.3.2.3. Matriz de Comprimentos de Corrida de Cinza - GLRLM

Dada uma imagem, define-se que um conjunto composto de pixels consecutivos, apresentando o mesmo nível de cinza e sendo colineares em uma dada direção, representa uma **corrida de cinza** (Figura 2.9). O número de pixels contidos nesse conjunto denota o comprimento da corrida.

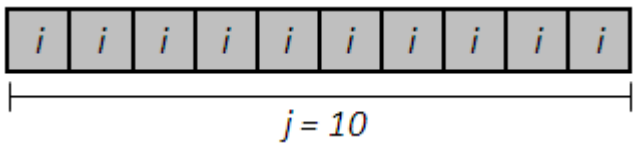


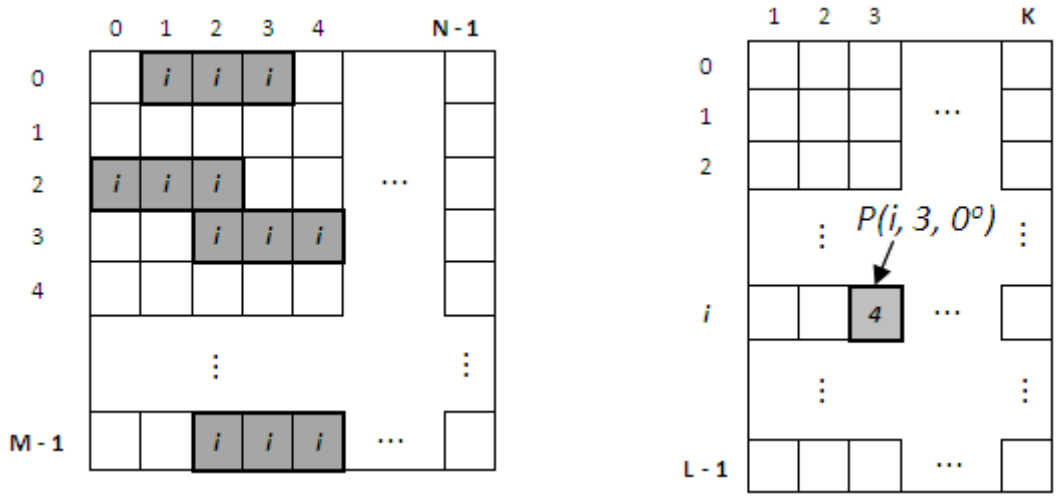
Figura 2.9: Exemplo de uma corrida de nível de cinza i, de comprimento 10 e direção horizontal.

Com o objetivo de sintetizar as informações obtidas a partir dessas corridas, são criadas matrizes de comprimentos de corrida de cinza (GLRLM, do inglês *Gray Level Run Length*

Matrix), onde cada elemento, representado por $P(i, j, \theta)$, contém o número de corridas com tamanho j (comprimento), tendo i como o nível de cinza de seus pixels, e o parâmetro θ como a orientação do segmento de reta formado pelos pixels (Figura 2.10). A partir da GLRLM podem ser extraídas medidas usadas para análise de textura (GALLOWAY, 1975). O cálculo do elemento da GLRLM (BEBIS, BOYLE, *et al.*, 2006) é definido como a seguir:

$$P(i,j,\theta) = CARD[\{(m,n)|f(m,n) = i, \tau(i,\theta) = j\}] \quad (9)$$

onde $f(m,n)$ denota a função nível de cinza no pixel (m,n). E $\tau(i,\theta)$ é o comprimento da corrida de nível de cinza i e direção θ, e $CARD$ significa a cardinalidade (número de elementos) do conjunto. Os valores de θ adotados são 0°, 45°, 90° e 135°.



(a) (b)

Figura 2.10: (a) Imagem de $M\ x\ N$ pixels. (b) Matriz de Comprimentos de Corrida de Cinza da imagem $(\theta = 0^o)$.

A Figura 2.10b ilustra a estrutura da GLRLM, construída a partir da imagem da Figura 2.10a. O tamanho da matriz é $L\ x\ K$, sendo L a quantidade máxima de níveis de cinza que a imagem pode apresentar e K o maior comprimento de corrida presente na imagem na direção θ. Na imagem (Figura 2.10a), por exemplo, há 4 corridas de nível de cinza i, comprimento 3 e direção horizontal. Assim, a célula $(i, 3)$ da GLRLM registra a frequência $P(i, 3, 0^o) = 4$.

2.3.2.4. Matriz de Comprimentos de Lacuna de Cinza - GLGLM

Dada uma imagem, define-se que uma **lacuna** (*gap*) para o nível de cinza g (XINLI, ALBREGTSEN e FOYN, 1994) ocorre quando g é encontrado apenas no início e no fim de um conjunto de pixels consecutivos e colineares, enquanto todos os valores de pixels entre estão acima ou abaixo de g (Figura 2.11). O comprimento da lacuna é a distância entre estes dois pixels menos um, de modo que, dois pixels vizinhos adjacentes com nível de cinza idêntico têm comprimento de lacuna zero. No caso onde nenhum pixel com nível de cinza g é encontrado ao longo da direção de busca, o comprimento da lacuna é considerado como infinito, sendo omitido.

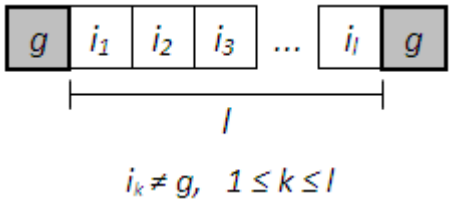
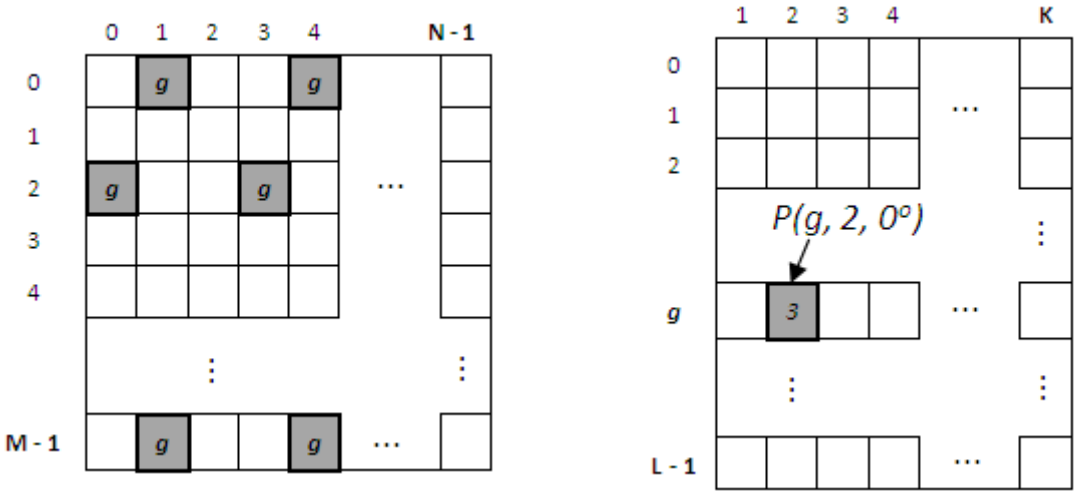


Figura 2.11: Lacuna de nível de cinza g, de comprimento l e direção horizontal.

A matriz de comprimentos de lacuna de nível de cinza (GLGLM, do inglês *Gray Level Gap Length Matrix*) é uma matriz estatística de ordem superior, onde cada elemento (g, l) armazena a frequência denotada por $P(g, l, \theta)$, com que lacunas de nível de cinza g, tamanho l, e inclinação θ ocorrem na imagem (Figura 2.12). O elemento da GLGLM (XINLI, ALBREGTSEN e FOYN, 1994), na direção θ, é definido como:

$$\begin{aligned} P(g, l, \theta) = \#\{(i,\ j) \mid\ & f(i,\ j) = g, \\ & f(i + x,\ j + y) = g, \\ & f(i + u,\ j + v) \neq g, \\ & x = (l + 1) \cdot \cos\theta, \\ & y = (l + 1) \cdot \operatorname{sen}\theta, \\ & u < x, v < y \} \end{aligned} \tag{10}$$

onde "#" denota o número de elementos do conjunto, e $f(i, j)$ denota a função nível de cinza no pixel (i, j).



(a) (b)

Figura 2.12: (a) Imagem de $M \; x \; N$ pixels. (b) Matriz de Comprimentos de Lacuna de Cinza da imagem $(\theta = 0^o)$.

A Figura 2.12b ilustra a estrutura da GLGLM, construída a partir da imagem da Figura 2.12a. O tamanho da GLGLM é $L \; x \; K$, sendo L a quantidade máxima de níveis de cinza que a imagem pode apresentar e K o maior comprimento de lacuna de cinza presente na imagem na direção θ. Na imagem (Figura 2.12a), por exemplo, há 3 lacunas de nível de cinza g, comprimento 2 e inclinação horizontal. Assim, a célula $(g, 2)$ da GLGLM registra a frequência $P(g, 2, 0^o) = 3$.

2.3.3 Equalização de histograma

A equalização de histograma é um método que modifica o histograma da imagem original f de tal forma que a imagem transformada g possua uma distribuição mais uniforme dos seus níveis de cinza, de modo que estes apareçam na imagem aproximadamente com a mesma frequência (PEDRINI e SCHWARTZ, 2008). Esta técnica é útil para realçar o contraste[3] da imagem (Figura 2.13a-c), uma vez que aumenta a escala dinâmica dos níveis de cinza (Figura 2.13b-d). Segundo (GONZALEZ e WOODS, 2000), a equalização de histograma possui a vantagem de ser completamente automática com relação às técnicas manuais de alteração de contraste. Entretanto, de acordo com (PEDRINI e SCHWARTZ,

[3] Corresponde à diferença entre os tons de cinza. Desta forma, baixo contraste ocorre quando há pequena diferença entre os níveis de cinza dos pixels localizados em uma região contígua da imagem (PEDRINI e SCHWARTZ, 2008).

2008), em algumas situações a equalização pode degradar uma imagem, como, por exemplo, uma imagem com grande concentração de pixels em poucos níveis de cinza.

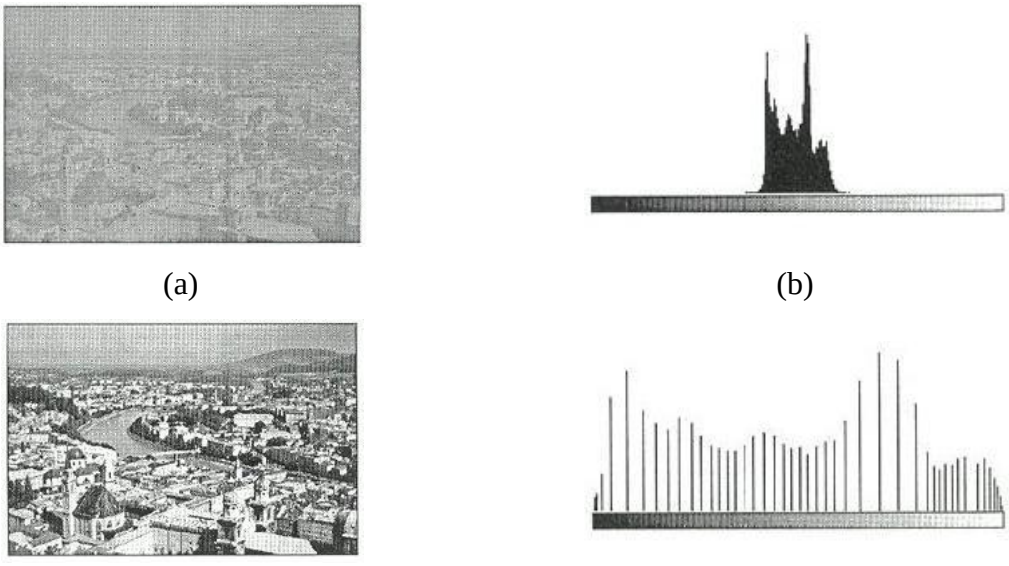
O processo de equalização, adotado neste trabalho, compreende o cálculo do histrograma da imagem, da função de distribuição acumulada (histograma acumulado) e do histograma normalizado, da seguinte forma: seja a imagem S, de dimensões M e N, com níveis de cinza no intervalo $[0, L-1]$, e $h(r_k)$ o histograma da imagem S. Desta forma, o histograma acumulado de $h(r_k)$, denotado por $H(r_k)$, é calculado como:

$$H(r_k) = H(r_k - 1) + h(r_k) \tag{11}$$

onde $H(0) = h(0)$ e $r_k = 1, ..., L-1$. Assim, o histograma normalizado, denotado por $T(r_k)$, é obtido a partir de $H(r_k)$, da seguinte forma:

$$T(r_k) = round\left(\frac{L-1}{MN} \cdot H(r_k) \right) \tag{12}$$

Portanto, a k-ésima intensidade, i_k, da imagem equalizada, pode ser obtido por $i_k = T(r_k)$. Na Figura 2.13c podemos observar o efeito da equalização sobre a imagem original (Figura 2.13a). Na mesma figura são exibidos o histograma da imagem original e o histograma da imagem equalizada.



(a) (b)

(c) (d)

Figura 2.13: Equalização do histograma. (a) Imagem original, (b) Histograma da imagem original, (c) Imagem equalizada e (d) Histograma da imagem equalizada. Adaptado de (MATARREDONA, 1994)

2.4 Reconhecimento de Padrões

Um padrão, segundo (LOONEY, 1997), é tudo aquilo para o qual existe uma entidade nomeável representante, geralmente, criada através do conhecimento cultural humano. O reconhecimento de padrões visa determinar um mapeamento que relacione as propriedades extraídas de amostras com um conjunto de rótulos (entidade nomeável representante), apresentando a restrição de que amostras com características semelhantes devem ser mapeadas ao mesmo rótulo. Os algoritmos que estabelecem este mapeamento são denotados como algoritmos de classificação ou classificadores (PEDRINI e SCHWARTZ, 2008).

Quando o processo de classificação considera classes previamente definidas, este é denominado como classificação supervisionada. Para que os parâmetros que caracterizam cada classe sejam obtidos, uma etapa denominada treinamento deve ser executada anteriormente à aplicação do algoritmo de classificação. Tais parâmetros são obtidos a partir de amostras previamente definidas (rotuladas). Quando não se dispõe de parâmetros ou informações coletadas previamente à aplicação do algoritmo de classificação, o processo é denominado como não-supervisionado. Neste caso, todas as informações de interesse devem ser obtidas a partir das próprias amostras a serem rotuladas (PEDRINI e SCHWARTZ, 2008). Neste trabalho utiliza-se a classificação supervisionada.

As medidas resultantes do processo de extração de características são representadas por meio de um vetor, x_i, denominado vetor de características (*feature vector*), composto de n elementos, na forma:

$$x_i = [x_{i1}\ x_{i2}\ \ldots x_{ij}\ \ldots x_{in}] \tag{13}$$

onde cada elemento, x_{ij}, representa uma das medidas utilizadas para descrever as propriedades (*features*) da amostra. Uma vez que os vetores de características são construídos, um para cada amostra, um algoritmo de classificação deve ser aplicado com o objetivo de atribuir as amostras às classes consideradas no experimento. No caso da classificação supervisionada, dado um conjunto de amostras rotuladas, na forma (x_i, y_i), em que x_i representa o vetor de características de uma amostra e y_i denota o seu rótulo, deve-se produzir, na etapa de treinamento, um classificador, também denominado modelo, preditor ou hipótese, capaz de predizer precisamente o rótulo de novas amostras, não conhecidas na etapa de treinamento (Figura 2.14). O modelo obtido pode ser encarado, também, como uma função

f, a qual recebe uma amostra x e fornece uma predição (classe) y (LORENA e CARVALHO, 2007).

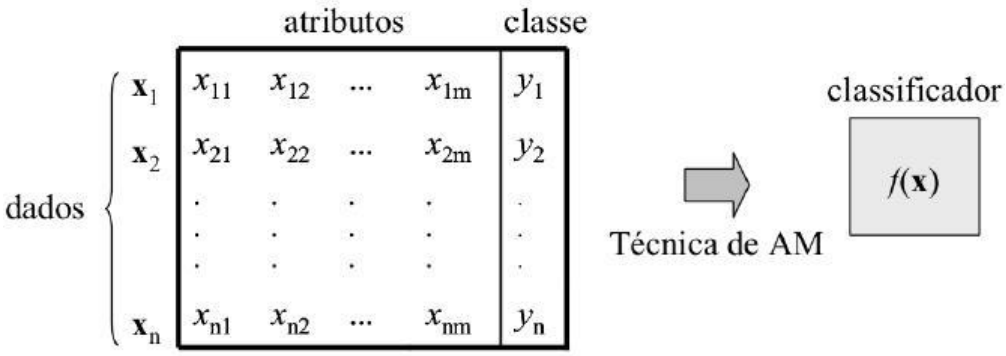
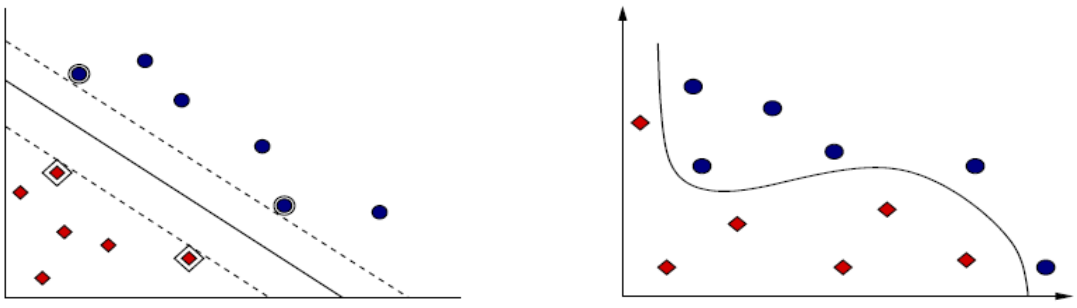


Figura 2.14: Classificação supervisionada. Fonte: (LORENA e CARVALHO, 2007)

O processo decisório, que determina qual classe uma amostra pertence, em problemas de reconhecimento de padrões pode ser realizado através de superfícies de decisão (funções) que dividem o espaço de características em regiões (SANTOS, 2002). De acordo com a superfície de decisão, os algoritmos de classificação podem ser separados em dois grupos: lineares e não-lineares (Figura 2.15). Os classificadores lineares são aqueles em que a superfície de decisão é uma reta ou um hiperplano[4] quando o vetor de características contém mais de duas medidas. Por outro lado, quando a superfície de decisão é mais complexa que uma reta ou hiperplano, o classificador é denominado não-linear (PEDRINI e SCHWARTZ, 2008).



(a) (b)

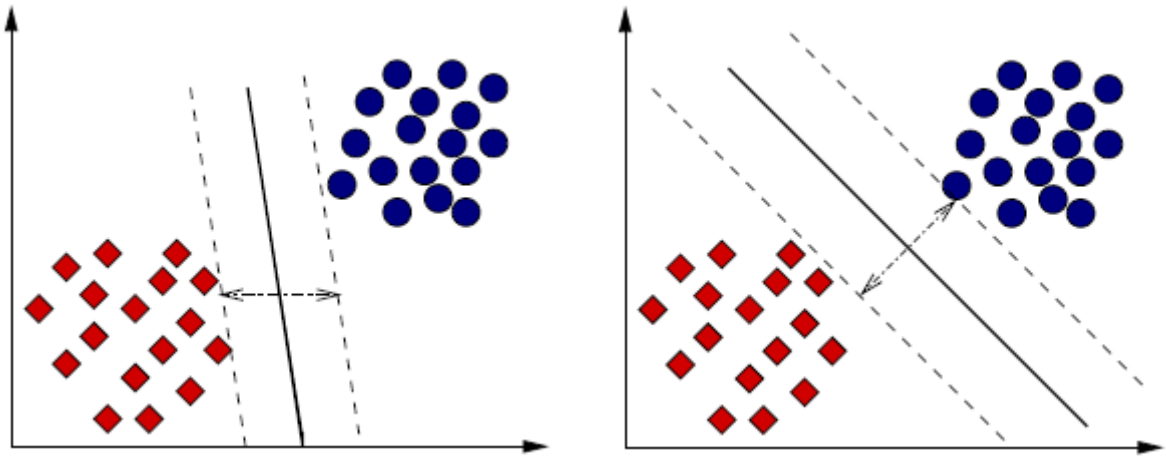
Figura 2.15: Superfícies de decisão. (a) linerares e (b) não-linerares. Fonte: (SANTOS, 2002)

[4] Generalização do *plano* em diferentes números de dimensões.

Neste trabalho utiliza-se a Máquina de Vetores de Suporte como o classificador supervisionado para realizar o reconhecimento de padrões de tecidos da mama, de modo a classificá-los em massa ou não-massa.

2.4.1 Máquina de Vetores de Suporte

A Máquina de Vetores de Suporte (SVM, do inglês *Support Vector Machine*) é uma técnica de aprendizagem supervisionada, desenvolvida por Vladmir Vapnik e colaboradores (VAPNIK, 1998), cujo princípio básico é a construção de um hiperplano como superfície de decisão, cuja margem de separação entre as classes seja máxima (Figura 2.16b). A margem é definida como a distância entre os pontos de dados, de ambas as classes, mais próximos ao hiperplano (SANTOS, 2002).



(a) (b)

Figura 2.16: (a) Um hiperplano de separação com margem pequena. (b) Um hiperplano de margem máxima. Fonte: (SANTOS, 2002)

A SVM tem se apresentado como um classificador superior, com grande habilidade de generalização[5], quando comparado a outros classificadores em uma variedade de aplicações (CRISTIANNI e SHAWE-TAYLOR, 2000). Destaca-se, também, por apresentar uma sólida fundamentação teórica e por ser útil em problemas com um grande número de entradas (SANTOS, 2002). Alguns exemplos de aplicações bem-sucedidas podem ser encontrados em diversos domínios, como a categorização de textos, na análise de imagens e em bioinformática (LORENA e CARVALHO, 2007).

[5] Capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu.

O modelo mais simples de SVM, e o primeiro a ser introduzido, chamado **Classificador de Margem Máxima**, trabalha apenas com dados linearmente separáveis, ficando, desta forma, restrito à poucas aplicações práticas (SANTOS, 2002). No contexto de classificação binária[6], este classificador consegue encontrar um hiperplano de separação ótima, que separa um conjunto de vetores sem erro e a distância entre os vetores (das classes opostas) mais próximos ao hiperplano, conhecidos como **vetores de suporte**, é máxima (VAPNIK, 1998). Os vetores de suporte representam os elementos críticos do conjunto de treinamento, de modo que, se forem removidos devem alterar a solução encontrada (SANTOS, 2002). De acordo com (BURGES, 1998) se os demais pontos (vetores) forem removidos e o treinamento for repetido, o mesmo hiperplano deve ser encontrado. Desta forma, os vetores de suporte são os únicos envolvidos na construção do Hiperplano de Margem Máxima (SANTOS, 2002).

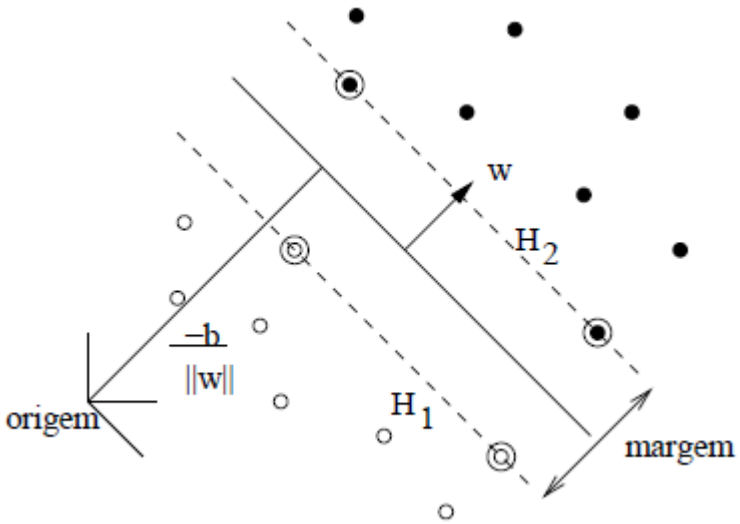


Figura 2.17: Hiperplano de separação para o caso linearmente separável. Os vetores de suporte estão circulados. Fonte: (SANTOS, 2002)

Quando o classificador de margem máxima é aplicado a dados não linearmente separáveis, não encontra a solução desejada, uma vez que este classificador constrói hipóteses baseadas na inexistência de erros de treinamento. Para manipular dados não linearmente separáveis o SVM utiliza o **Hiperplano de Margem Suave** (*soft margin*), que permite tolerar ruídos[7] e *outliers*[8], considerando mais pontos de treinamento, além dos que estão na fronteira (Figura 2.18) e permitindo a ocorrência de erros de classificação. Para minimizar os erros de

[6] O conjunto de amostras é constituído por duas classes.
[7] Representam casos onde os vetores de características apresentam rótulo e/ou atributos incorretos.
[8] Pontos muito distantes das classes a que pertencem.

classificação, são introduzidos variáveis de folga, que representa o custo extra para os erros, permitindo a classificação correta de *outliers*, e um parâmetro C, definido pelo usuário (SANTOS, 2002). Este parâmetro é uma constante que atua como uma função de penalidade, prevenindo que *outliers* afetem o hiperplano ótimo (KWONG e GONG, 1999), estabelecendo um *trade-off* entre as violações do hiperplano (*outliers*) e o tamanho da margem (NOBLE, 2006). De acordo com (NOBLE, 2006), a definição deste parâmetro é complicado pelo fato de se tentar alcançar uma margem grande com respeito aos exemplos classificados corretamente.

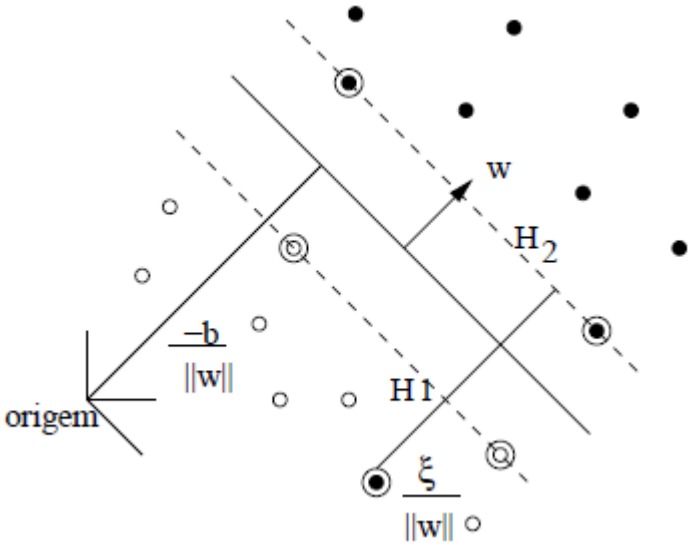


Figura 2.18: Hiperplano de separação para o caso não linearmente separável. Fonte: (SANTOS, 2002)

Na Figura 2.18, H_1 e H_2 representam as fronteiras das classes, o símbolo ξ representa uma variável de folga, e o ponto preto circulado, referenciado pelo símbolo ξ, um *outlier*.

Outra forma, mais robusta, utilizada pelo SVM para permitir a classificação de dados não linearmente separáveis, é com a aplicação do conceito de **Funções Kernel**, cuja idéia é projetar dados a partir de um espaço de baixa dimensão para um espaço de alta dimensão. A estratégia, neste caso, envolve mudar a representação dos dados da seguinte forma:

$$x = (x_1, ..., x_n) \quad \mapsto \quad \phi(x) = (\phi_1(x), ..., \phi_N(x))$$

que corresponde ao mapeamento do espaço de entrada X, em que os dados são não linearmente separáveis, em um novo espaço $Z = \{\phi(x) \mid x \in X\}$, chamado espaço de características, de dimensão maior, onde os dados passam a ser linearmente separáveis, em que ϕ_i são as funções kernel. Embora a dimensão do espaço aumente em Z, a complexidade

diminui, porque a classificação, que no espaço de entrada só era possível utilizando superfícies de decisão não lineares, no espaço de características, pode ser feita apenas com um simples hiperplano (Figura 2.19). Neste caso, o hiperplano de separação é definido como uma função linear de vetores do espaço de características e não do espaço de entrada original (SANTOS, 2002).

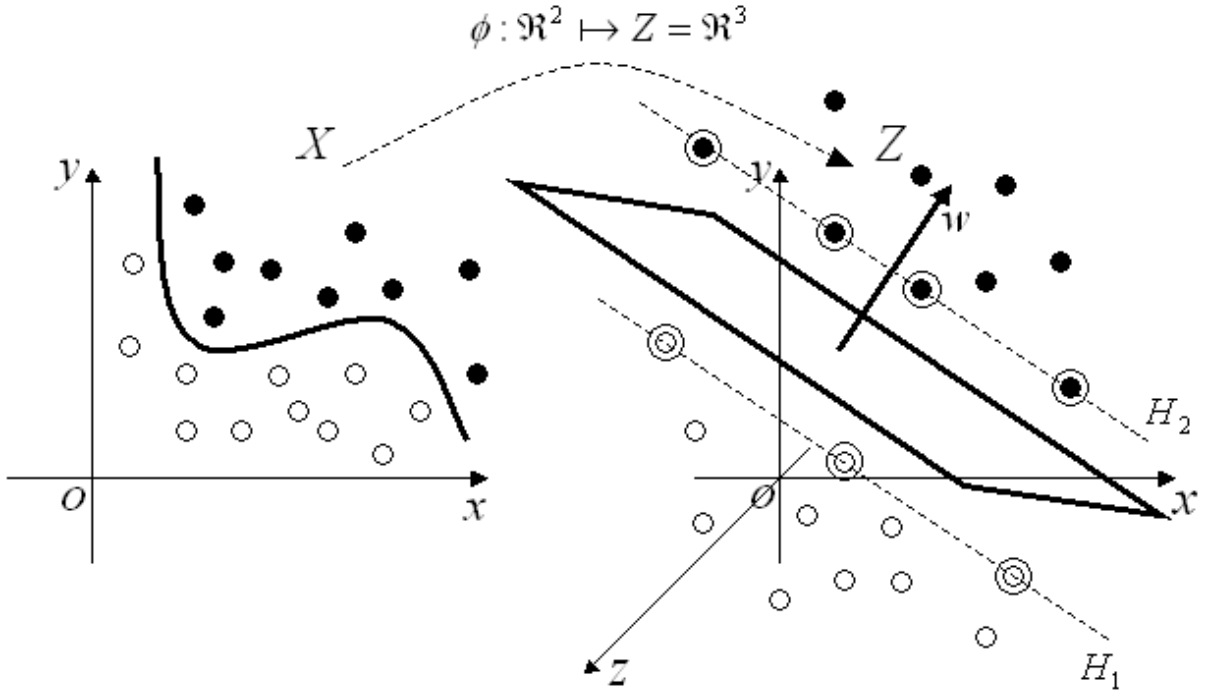


Figura 2.19: Mapeamento do espaço de entrada para o espaço de características via função kernel.

A Figura 2.19 ilustra um exemplo de projeção de dados, não linearmente separáveis, de um espaço de entrada bi-dimensional, $X = \Re^2$, para o espaço de características tri-dimensional, $Z = \Re^3$, onde tornam-se linearmente separáveis. As funções kernel mais usadas são a Polinomial, Função de Base Radial Gaussiana (RBF) e a Rede Neural Sigmóide (SANTOS, 2002). Neste trabalho utiliza-se a função de base radial, que é definida da seguinte forma:

$$K(x, x_i) = e^{-\gamma \|x - x_i\|^2} \tag{14}$$

2.5 Métricas de Validação de Resultados

Na análise de imagens médicas, geralmente utiliza-se algumas estatísticas descritivas sobre os resultados dos testes para avaliar o desempenho do classificador, como sensibilidade (S), especificidade (E) e acurácia (A) (BLAND, 2000). Estas métricas são calculadas a partir de quatro situações possíveis em relação ao diagnóstico:

- VP – Verdadeiro Positivo: o teste é positivo e o paciênte têm a doença;
- FP – Falso Positivo: o teste é positivo, mas o paciênte não têm a doença;
- VN – Verdadeiro Negativo: o teste é negativo e o paciênte não têm a doença;
- FN – Falso Negativo: o teste é negativo, mas o paciênte têm a doença.

A **sensibilidade** define a proporção de verdadeiros-positivos identificados no teste. Indica quão bom é o teste para identificar indivíduos doentes:

$$S = \frac{VP}{VP + FN} \tag{15}$$

A **especificidade** define a proporção de verdadeiros-negativos identificados no teste. Indica quão bom é o teste para identificar indivíduos não doentes:

$$E = \frac{VN}{VN + FP} \tag{16}$$

A **acurácia** define a taxa de casos identificados corretamente sobre o número total de casos:

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{17}$$

Neste trabalho, utilizamos estas três medidas para avaliar o desempenho da metodologia proposta.

3 METODOLOGIA

Neste capítulo são descritos os procedimentos utilizados na metodologia proposta para classificação de regiões de tecidos da mama, extraídos a partir de imagens mamográficas, em massa e não massa. Inicialmente, é apresentado a base de imagens utilizada nos experimentos. Em seguida, descrevemos os passos necessários para a extração de características das amostras, utilizando o índice de diversidade de McIntosh, e a posterior classificação das mesmas, através da submissão dos vetores de características obtidos ao classificador supervisionado SVM, finalizando a metodologia com a técnica de validação dos resultados.

3.1 Metodologia Proposta

A Figura 3.1 apresenta a sequência de etapas usadas na metodologia proposta neste trabalho.

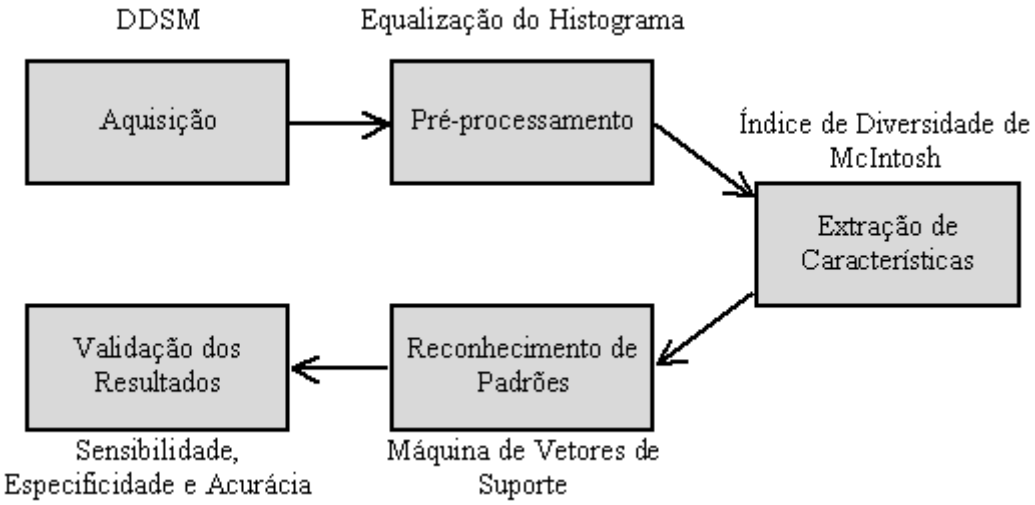


Figura 3.1: Etapas da metodologia proposta.

Na primeira etapa, chamada aquisição, são obtidas as amostras de mamografias normais e com regiões suspeitas, a partir das quais são extraídas regiões de interesse, de tecidos de massa e não-massa, necessárias para o processamento. A seguir, as regiões de interesse obtidas são submetidas a um pré-processamento, através da equalização de histograma, visando realçar os aspectos de textura presentes nas imagens. Na etapa seguinte, de extração

de características, são computados, para cada amostra, os valores do índice de diversidade de McIntosh, considerando-se diferentes quantizações.

A partir das características extraídas é gerado um vetor de características. Entramos, então, na etapa de reconhecimento de padrões, onde o conjunto destes vetores é submetido ao classificador supervisionado SVM. Nesta etapa, um subconjunto de vetores de características é utilizado na etapa de treinamento, onde é gerado um modelo (função) de classificação. O outro subconjunto é, então, classificado através deste modelo. Por fim, os resultados da classificação são usados na etapa de validação.

3.1.1 Aquisição das Amostras

Neste trabalho foram utilizadas amostras de mamografias digitalizadas do banco de dados DDSM - *Digital Database for Screening Mammography* (HEATH, BOWYER, *et al.*, 2000), o qual é disponibilizado gratuitamente na Web. A base DDSM possui 2620 exames de pacientes de diferentes origens étnicas e raciais. Cada exame contém duas imagens de cada mama, nas projeções médio-lateral oblíqua e crânio-caudal. Além disso, são disponibilizadas informações sobre a paciente, tal como a idade e a densidade da mama. Junto com as imagens que apresentam áreas suspeitas (massas) é fornecido um arquivo de descrição de lesão (*overlay*), contendo a quantidade de lesões presentes na mamografia, a localização da lesão, o tipo de lesão, o contorno da lesão e seu diagnóstico. O contorno da lesão está codificado em *chain code* (MORSE, 2000).

Foram utilizadas 1850 regiões de interesse de tecidos normais (não-massa) e 1850 regiões de interesse de tecidos com massas (neoplasias benignas e malignas), totalizando 3700 amostras. Estas amostras provêm da mesma base DDSM do trabalho de (JUNIOR, 2008), onde o mesmo havia desenvolvido uma abordagem de extração automática das regiões de massa da seguinte forma: as amostras de tecidos de massas foram extraídas a partir do contorno da lesão, através da aplicação de uma *bounding box*[9] sobre o contorno. Os pixels entre o contorno e a *bounding box* tiveram suas intensidades setadas para -1, por representarem áreas que não constituem massa, e que podem influenciar no processo de extração de características. A nova imagem foi armazenada em um arquivo de texto, onde cada linha contém a coordenada espacial (x, y) do pixel e seu nível de cinza. Desta forma, apenas os pixels da região interna ao contorno (valor diferente de "–1") são considerados nos

[9] Menor retângulo que engloba uma região.

processos seguintes, de pré-processamento e extração de características. As amostras de não-massas foram retiradas, de forma aleatória e manual, de mamografias sem suspeita de anormalidades.

(a) (b)

Figura 3.2: Regiões extraídas de mamografias da base DDSM. (a) massas, (b) não massas.

Todas as regiões de interesse extraídas de acordo com esta abordagem apresentaram tamanhos diferentes, uma vez que, para conservar o máximo de informação de textura presente nos tecidos de massas, era necessário conservar seus contornos, que variavam em forma e tamanho (Figura 3.2a). No entanto, isto não prejudicou o processo de extração de características, umas vez que, segundo (MELO, 2008), alguns índices de diversidade apresentam a vantagem de serem relativamente independentes do esforço amostral[10]. Com amostras relativamente pequenas podemos obter um valor de diversidade que mudará pouco conforme aumentamos o esforço amostral, de modo que podemos comparar diretamente comunidades estudadas com diferentes esforços amostrais (LLOYD, INGER e KING, 1968) (MAGURRAN, 2004).

3.1.2 Pré-Processamento

Após a aquisição das amostras, as mesmas foram submetidas ao processo de pré-processamento, por meio da equalização de histograma (Seção 2.3.3). O uso da equalização foi considerado por causa que a sua aplicação gera uma redistribuição mais homogênea dos níveis de cinza, ocasionando um realce da textura do tecido, uma vez que alguns aspectos dissimilares de textura, que na amostra original podem ficar ocultos, apresentando, desta forma, pouca contribuição na discriminação, passam a ter maior relevância. O índice de diversidade, desta forma, tende a ser mais representativo para descrever a textura do tecido, aumentando, assim, a capacidade da acurácia geral desta metodologia.

[10] O esforço amostral refere-se à área coletada, ou ao tempo de coleta, ou ao número de indivíduos coletados.

(a) (b)

Figura 3.3: (a) Massa original, (b) Massa após a equalização do histograma.

3.1.3 Extração de Características

Na fase de extração de características, inicialmente, aplicamos a técnica de quantização uniforme (Seção 2.3.1), com o objetivo de agregar as informações de textura presentes em cada quantização e, assim, aumentar o poder discriminatório. As amostras foram quantizadas em 256, 128, 64, 32, 16 e 8 níveis de cinza. A partir de cada quantização, é calculado um índice de diversidade de McIntosh (Seção 2.2.1), para descrever a textura da amostra. Este cálculo é proposto através de quatro abordagens independentes: (1) a partir do histograma da imagem; (2) a partir da Matriz de Co-ocorrência de Níveis de Cinza (GLCM); (3) a partir da Matriz de Comprimentos de Corrida de Cinza (GLRLM); e (4) a partir da Matriz de Comprimentos de Lacuna de Cinza (GLGLM).

3.1.3.1. Índice de Diversidade de McIntosh a partir do Histograma

A idéia, nesta abordagem, já usada em trabalhos como (SILVA, 2009), com o índice de diversidade de simpson e (SOUSA, 2011), com o índice de diversidade de Shannon, é calcular a diversidade de níveis de cinza presentes na imagem, sem levar em consideração as relações espaciais entre pixels vizinhos, e usá-la como um atributo de textura. Desta forma, a entidade espécie é definida como sendo o nível de cinza (Figura 3.4a). Como o histograma (Seção 2.3.2.1) registra a frequência de cada nível de cinza (espécie) da imagem, a partir dele podemos extrair a **riqueza de espécies** (s), representada pela quantidade de entradas não nulas (*bins*) do histograma, e a **abundância relativa** de cada espécie, representada pelo valor de cada *bin*. Assim, os parâmetros s, N e U, necessários para o cálculo do índice de diversidade de McIntosh (Equação 3), são obtidos da seguinte forma:

$$s = \#\{H(i) \mid H(i) \neq 0, 0 \leq i < L\},$$

$$N = \sum_{k=1}^{s} H(i_k) \quad \text{e} \quad U = \sqrt{\sum_{k=1}^{s} (H(i_k))^2}$$

onde "#" significa o número de elementos do conjunto e $H(i)$ denota a entrada do histrograma (frequência do nível de cinza i). Uma vez que calculamos um índice de diversidade para cada quantização, o vetor de características gerado apresentou 6 variáveis.

3.1.3.2. Índice de Diversidade de McIntosh a partir da GLCM

A proposta desta abordagem é calcular a diversidade de transições de nível de cinza entre dois pixels vizinhos, partindo da suposição de que em um tecido possa ocorrer, em geral, para algumas quantizações, mais transições[11] do que a quantidade de transições do outro tecido, o que seria útil como um fator discriminante. Desta forma, a entidade espécie é representada por um par de pixels, separados por uma distância d, e alinhados sob uma direção θ (Figura 3.4b). Assim, a GLCM (Seção 2.3.2.2) passa a representar a distribuição das espécies, de modo que podemos extrair a **riqueza de espécies** (s), representada pela quantidade de entradas não nulas da matriz, e a **abundância relativa** de cada espécie, representada pelo valor contido em cada entrada não nula. Considerando $P(i, j, d, \theta)$ o valor da entrada (i, j) da GLCM, então os valores dos parâmetros s, N e U, necessários para o cálculo do índice de diversidade de McIntosh (Equação 3) são obtidos da seguinte forma:

$$s = \#\{P(i, j, d, \theta) \mid P(i, j, d, \theta) \neq 0\},$$

$$N = \sum_{k=1}^{s} P(i_k, j_k, d, \theta) \quad \text{e} \quad U = \sqrt{\sum_{k=1}^{s} (P(i_k, j_k, d, \theta))^2}$$

Os valores adotados para a direção θ foram 0^o, 45^o, 90^o e 135^o, e para a distância d foram 1, 2, 3, 4 e 5. Verificamos, empiricamente[12], que estas 5 distâncias contribuiam com resultados mais expressivos para discriminação dos tecidos de massa e não massa. Desta forma, como era necessária uma GLCM para cada θ e d, e foram consideradas seis quantizações, o vetor de características gerado apresentou 120 atributos de textura (5 distâncias x 4 direções x 6 quantizações).

---

[11] Com muitas transições de nível de cinza, o tecido tende a apresentar uma textura mais fina. Com poucas transições, o tecido apresenta uma textura mais grossa, mais homogênea.

[12] Foram feitos testes com até 7 distâncias. Os melhores resultados foram constatados com 5 distâncias.

3.1.3.3. Índice de Diversidade de McIntosh a partir da GLRLM

De acordo com (PEDRINI e SCHWARTZ, 2008) em texturas ásperas, espera-se que corridas relativamente longas sejam frequentes, por outro lado, corridas mais curtas ocorrem em texturas finas. Assim, se um tecido apresentar muitas corridas longas e poucas corridas curtas (textura áspera), e outro apresentar muitas corridas curtas e poucas corridas longas (textura fina), então a variedade de corridas de cinza do primeiro tecido tende a ser menor que a do segundo, pois, para preencher uma área, as corridas longas acomodam-se em menor número que as corridas curtas, o que ocasiona uma distribuição menor de corridas. Desta forma, supondo que, em geral, o tecido de massa apresente textura fina e o tecido de não massa apresente textura grossa (áspera), ou vice-versa, então propomos o cálculo da diversidade de comprimentos de corrida de cinza como atributo de textura, de modo a viabilizar a discriminação entre estes tecidos. Para adaptar o conceito de diversidade ecológica, adotamos que a entidade espécie seja representada por uma corrida de cinza de intensidade i, comprimento j e inclinação θ (Figura 3.4c). Assim, a GLRLM (Seção 2.3.2.3) passa a representar a distribuição das espécies da região de interesse. Desta forma, a partir da GLRLM são extraídos a **riqueza de espécies**, representada pela quantidade de entradas não nulas da matriz, e a **abundância relativa** de cada espécie, representada pelo valor contido em cada uma destas entradas não nulas. Considerando $P(i, j, \theta)$ o valor da entrada (i, j) da GLRLM, para a direção θ, então os valores dos parâmetros s, N e U, necessários para o cálculo do índice de diversidade de McIntosh (Equação 3) são obtidos da seguinte forma:

$$s = \#\{P(i,\ j, \theta) \mid P(i,\ j, \theta) \neq 0\}$$

$$N = \sum_{k=1}^{s} P(i_k,\ j_k, \theta) \quad \text{e} \quad U = \sqrt{\sum_{k=1}^{s} (P(i_k,\ j_k, \theta))^2}$$

Como é necessário uma GLRLM para cada direção, utilizamos quatro GLRLM, adotando os valores de θ igual a 0^o, 45^o, 90^o e 135^o. Desta forma, para as seis quantizações consideradas, foi gerado um vetor de característica com 24 variáveis.

3.1.3.4. Índice de Diversidade de McIntosh a partir da GLGLM

Propomos, nesta abordagem, o cálculo da diversidade de lacunas de cinza presentes na imagem como atributo de textura. Assim, a entidade espécie é representada pela lacuna de nível de cinza g, comprimento l e inclinação θ (Figura 3.4d). Uma vez que os pixels extremos

de uma lacuna são homogêneos (mesmo nível de cinza), podemos encarar a lacuna de comprimento l como uma **vizinhança homogênea** de tamanho $l+1$. Desta forma, se um tecido apresentar, de modo geral, a textura mais homogênea que o outro, é provável que ele contenha uma concentração maior de vizinhanças homogêneas, o que sujere o índice de diversidade dessas vizinhanças homogêneas (ou lacunas) como um descritor de textura. Neste caso, a GLGLM (Seção 2.3.2.4) passa a representar a distribuição das espécies da região de interesse. Deste modo, a partir da GLGLM são obtidos a **riqueza de espécies** (s), representada pela quantidade de entradas não nulas da matriz, e suas **abundâncias relativa**, representada pelo valor contido em cada uma destas entradas não nulas. Assim, considerando $P(g,l,\theta)$ o valor da célula (g,l) da GLGLM, para a direção θ, então os valores dos parâmetros s, N e U, necessários para o cálculo do índice de diversidade de McIntosh (Equação 3) são obtidos da seguinte forma:

$$s = \#\{P(g,l,\theta) \mid P(g,l,\theta) \neq 0\}$$

$$N = \sum_{k=1}^{s} P(g_k, l_k, \theta) \quad \text{e} \quad U = \sqrt{\sum_{k=1}^{s} (P(g_k, l_k, \theta))^2}$$

Como é necessário uma GLGLM para cada direção, utilizamos quatro GLGLM, adotando os valores de θ igual a 0^o, 45^o, 90^o e 135^o. Desta forma, para as seis quantizações consideradas, foi gerado um vetor de característica com 24 variáveis.

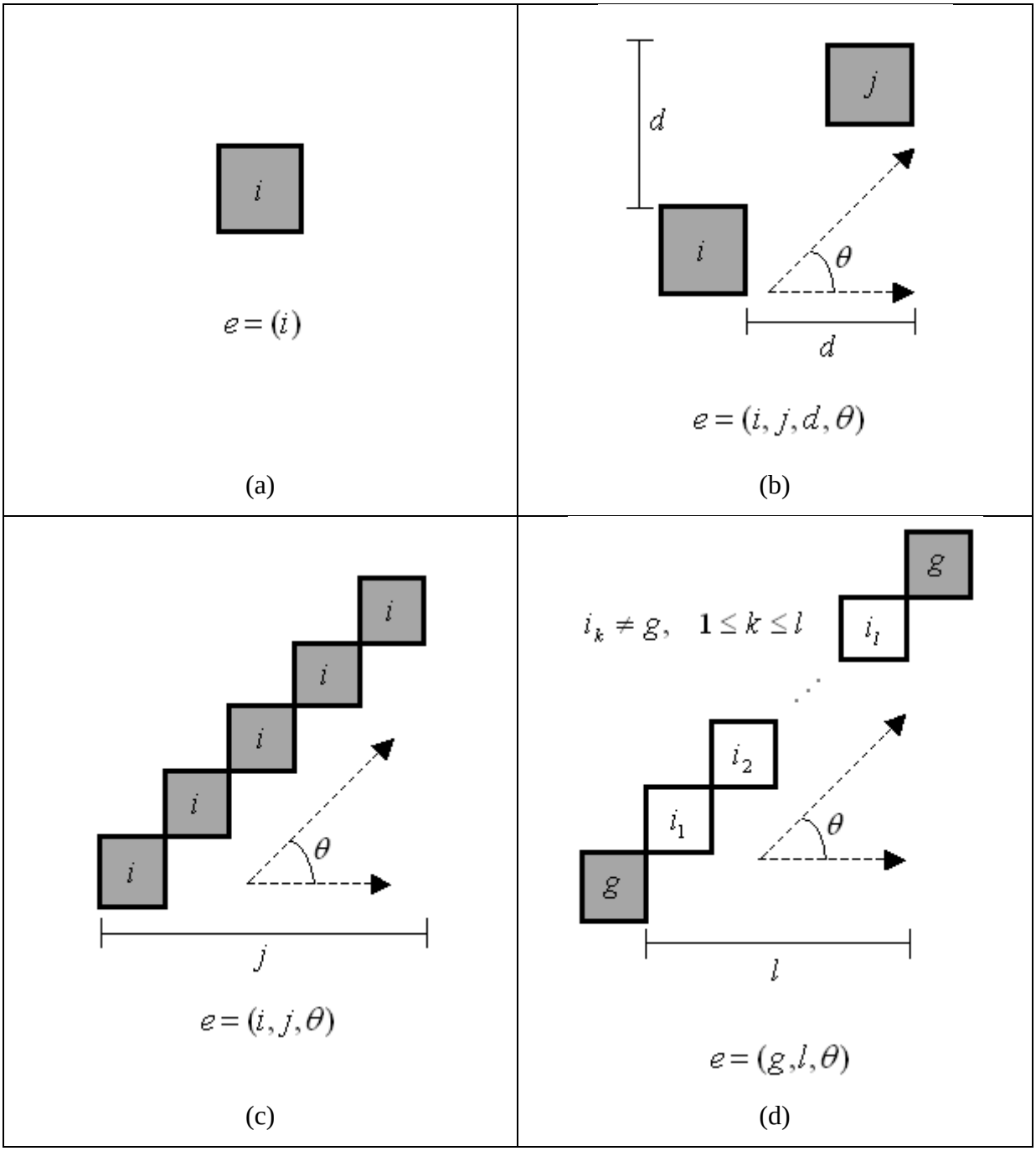


Figura 3.4: Definição da entidade espécie (*e*) na imagem, para cada abordagem. (a) Entidade espécie é o nível de cinza, (b) Entidade espécie é o par de pixel (transição de *i* para *j*), (c) Entidade espécie é a corrida de cinza de comprimento *j* e (d) Entidade espécie é a lacuna de cinza de tamanho *l*.

3.1.4 Reconhecimento de Padrões

A etapa final da metodologia proposta consiste em classificar as amostras em massa e não-massa, utilizando reconhecimento de padrões (Seção 2.4). Os vetores de características das amostras, gerados na etapa de extração de características, através do cálculo do índice de

diversidade de McIntosh, a partir das quatro abordagens propostas (histograma, GLCM, GLRLM e GLGLM), são submetidos ao classificador supervisionado SVM.

Durante a etapa de extração de características, é gerado uma base de características, onde cada linha é formada pelo rótulo (classe) de uma amostra, obtida da base DDSM, seguido pelo seu vetor de características. Foram geradas duas bases de características para cada abordagem: (1) para amostras equalizadas; (2) para amostras não equalizadas. A intenção foi comparar, em cada abordagem, os resultados dos experimentos[13], para avaliar a influência da equalização de histograma na discriminação das amostras.

Cada base de características foi dividida em dois grupos: base de treino e base de teste. O critério de divisão adotado foi 50% para treino e 50% para teste, sendo a divisão repetida, de forma randômica, 5 vezes. A intenção, neste critério, é verificar se as acurácias (acertos), em todas as repetições, comportam-se de modo semelhante, com valores altos e com pequena diferença entre eles, demonstrando, assim, que a abordagem testada representa bem o padrão de textura das amostras de massa e não massa. De acordo com (JUNIOR, 2008), é recomendável normalizar a base de características para uma faixa de valores comuns, como –1 à 1, para ajudar o classificador a convergir com maior facilidade na etapa de treinamento. No entanto, como o índice de diversidade de McIntosh (Equação 3) fornece valores entre 0 e 1, não foi necessário executar uma etapa extra de normalização.

Neste trabalho foi utilizado a função de base radial (RBF) (Seção 2.4.1), uma vez que o seu emprego em trabalhos relacionados à análise de imagens mamográficas, como em (JUNIOR, 2008) e (SOUSA, 2011), apresentou os melhores resultados. Sua utilização requer a estimação, com base nas amostras de treinamento, dos valores de dois parâmtros: C e γ. O parâmetro C define um peso dado aos erros de classificação das amostras de treinamento, de modo a minimizar a ocorrência de violações do hiperplano (*outliers*). O parâmetro γ, estimado para a função radial, é útil para otimizá-la, de modo que o SVM apresente a melhor eficácia para cada problema. Os valores destes parâmetros são estimados através de uma busca exaustiva realizada pelo script *grid.py*, em phyton, pertencente ao pacote LIBSVM[14] (CHANG e LIN, 2010). Este script busca, através de validação cruzada, a melhor combinação de parâmetros para a base, retornando o melhor percentual de acerto total (acurácia) sobre as amostras de treino e teste.

[13] O experimento compreende o treinamento e teste sobre as amostras de uma base de característica particular.
[14] Este pacote é a implementação SVM utilizada neste trabalho.

Durante a etapa de treinamento é gerado o modelo (função) que o SVM utiliza para classificar as amostras de teste. A etapa de treinamento desconhece as amostras de teste. Assim, o mecanismo de classificação busca se assemelhar com condições reais de teste. Durante a classificação, o SVM gera uma base de predições, contendo apenas os rótulos (classes) das amostras de teste.

3.1.5 Validação de Resultados

Com a etapa de classificação das amostras finalizada, é necessário validar os resultados e discutir possíveis melhorias. Desta forma, essa metodologia usa métricas comumente empregadas em sistemas CAD/CADx e aceitas pela sociedade para a análise de desempenho de sistemas baseados em processamento de imagens. Estas métricas são sensibilidade, especificidade e acurácia (Seção 2.5).

O processo de validação foi realizado da seguinte forma: para cada abordagem de extração de características, foram feitas comparações, linha a linha, dos rótulos (classes) da base de predições com os rótulos pré-definidos da base de características das amostras de teste, sendo que o resultado de cada comparação assume um dos valores: VN, FP, FN e VP (Seção 2.5). Uma vez contabilizados cada um destes valores, em toda a base de teste, é possível calcular as métricas de validação para cada abordagem.

A utilização das métricas de validação permite medir o desempenho da metodologia como satisfatória ou não, sendo útil, inclusive, para apontar pontos positivos e negativos para melhoria futura deste trabalho.

4 RESULTADOS E DISCUSSÃO

Este capítulo apresenta e discute os resultados obtidos com a metodologia proposta por este trabalho para classificação de regiões extraídas de imagens mamográficas em massa e não massa. São discutidos, inicialmente, os resultados individuais de cada uma das abordagens de extração de características apresentadas na Seção 3.1.3, baseadas no cálculo do Índice de Diversidade de McIntosh a partir do: (1) histograma, (2) GLCM, (3) GLRLM e (4) GLGLM. Além dessas 4, são consideradas mais duas abordagens: (5) onde o descritor de textura resulta da junção das diversidades calculadas a partir das GLRLMs com as diversidades calculadas a partir das GLCMs (GLRLM e GLCM), e (6) onde o descritor resulta da junção das diversidades calculadas a partir das GLRLMs com as diversidades calculadas a partir das GLGLMs (GLRLM e GLGLM), totalizando 6 abordagens. Com relação à etapa de treinamento, são apresentados os valores dos parâmetros C e γ, utilizados no núcleo radial (RBF). Estes parâmetros, conforme apresentado na Seção 3.1.4, são estimados automaticamente para cada conjunto de amostras de treinamento, e são utilizados durante a etapa de classificação pela Máquina de Vetores de Suporte. Os resultados da classificação são, então, avaliados através das métricas de validação (Seção 3.1.5). Finalizamos este capítulo com uma discussão sobre os melhores resultados obtidos em cada abordagem.

4.1 Resultados Obtidos

Esta seção apresenta e discute os resultados de cada abordagem. No intuito de verificar, em cada abordagem, a necessidade do pré-processamento, através da equalização de histograma, são apresentados os resultados de experimentos feitos com amostras equalizadas e não equalizadas. Sobre as abordagens que apresentaram melhores resultados foi aplicado o método *stepwise*, da Análise Discriminante Linear (ADL). Este método é útil para selecionar as variáveis independentes que melhor discriminam as classes, gerando um conjunto reduzido de variáveis para o modelo. O conjunto reduzido é geralmente melhor do que o conjunto completo de variáveis (HAIR, ANDERSON, *et al.*, 2005). No entanto, em todas as abordagens propostas neste trabalho, o uso do conjunto reduzido de variáveis gerou resultados menores que o uso de todas as variáveis. Desta forma, as discussões apresentadas nesta seção

dão ênfase aos resultados obtidos com todas as variáveis, sendo os resultados obtidos com os conjuntos reduzidos apenas brevemente citados.

4.1.1 Abordagem usando Histograma

A abordagem de extração de características usando o Índice de Diversidade de McIntosh calculado a partir do Histograma (Seção 3.1.3.1) gera um vetor composto de 6 características (um para cada quantização).

Para utilizar o classificador supervisionado SVM, foi necessário dividir a base de características em dois subconjuntos: base de treinamento e base de teste. Essa divisão foi repetida 5 vezes, na proporção 50% para treino e 50% para teste, através da seleção randômica[15] dos vetores de características da base. Em seguida, para usar a função de base radial (RBF) no classificador SVM, foi necessário estimar os valores dos parâmetros C e γ, aplicando o utilitário *grid.py* sobre as bases de treinamento. Estes valores são apresentados na Tabela 4.1.

Tabela 4.1: Parâmetros usados pelo classificador SVM, obtidos a partir das características extraídas com o Índice de Diversidade de McIntosh calculado a partir do histograma

	Conjunto de Treinamento	C	γ
Amostras não equalizadas	1	32768,00	2,0
	2	32768,00	8,0
	3	32768,00	8,0
	4	32768,00	2,0
	5	2048,00	8,0
Amostras equalizadas	1	32768,00	8,0
	2	2048,00	8,0
	3	32768,00	8,0
	4	512,00	8,0
	5	32768,00	8,0

Seguimos, então, com a etapa de classificação e validação dos resultados. Aplicando os parâmetros C e γ, acima, no classificador SVM, e as métricas de validação, obtemos os resultados apresentados na Tabela 4.2

[15] Foi utilizado o aplicativo *subset.py* do pacote LIBSVM.

Tabela 4.2: Resultados obtidos na classificação de amostras em massa e não massa utilizando o Índice de Diversidade de McIntosh a partir do Histograma.

	Conjunto de Teste	VP	FP	VN	FN	Sensibilidade	Especificidade	Acurácia
Amostras não equalizadas	1	505	213	698	434	53,78%	76,62%	65,03%
	2	603	318	590	339	64,01%	64,98%	64,49%
	3	***611***	***317***	***593***	***329***	***65,00%***	***65,16%***	***65,08%***
	4	570	326	625	329	63,40%	65,72%	64,59%
	5	613	328	584	325	65,35%	64,04%	64,70%
Amostras equalizadas	1	519	254	641	436	54,35%	71,62%	62,70%
	2	567	334	582	367	60,71%	63,54%	62,11%
	3	***621***	***418***	***529***	***282***	***68,77%***	***55,86%***	***62,16%***
	4	543	306	606	395	57,89%	66,45%	62,11%
	5	618	450	507	275	69,20%	52,98%	60,81%

Com as amostras não equalizadas, o melhor resultado obtido alcança a acurácia de 65,08%, sensibilidade de 65,00% e especificidade de 65,16%. A acurácia média é de 64,78%. Com as amostras equalizadas, os dois melhores resultados apresentam 62,70% e 62,16% de acurácia, porém o segundo resultado, embora com acurácia um pouco menor, é considerado melhor por apresentar uma sensibilidade consideravelmente maior, de 68,77% contra 54,35% do primeiro. Como a acurácia média, para estas amostras (equalizadas), foi de 61,98%, o que representa uma queda de 3% em relação aos resultados com amostras não equalizadas, conclui-se que a equalização de histograma, neste caso, prejudica a discriminação.

Em geral, tanto para amostras equalizadas quanto não-equalizadas, o desempenho da abordagem mostrou-se ruim. Provavelmente, este fenômeno esteja ocorrendo de acordo com a afirmação de Jahne (JAHNE, 2005), de que na estatística de primeira ordem (histograma), texturas com diferentes arranjos espaciais, mas com a mesma distribuição de valores de cinza não podem ser distinguidas, o que pode explicar a baixa acurácia atingida, de 65,08%. Outro problema, que podemos destacar nesta abordagem, é a taxa de falsos-positivos, que é muito alta, o que pode ocasionar muitas biópsias desnecessárias.

4.1.2 Abordagem usando GLCM

A abordagem de extração de características usando o Índice de Diversidade de McIntosh calculado a partir da GLCM (Seção 3.1.3.2) gera um vetor composto de 120 características (6 quantizações x 4 direções x 5 distâncias).

A base de características, da mesma forma que a abordagem anterior, foi dividida em dois subconjuntos, para utilização no classificador SVM: base de treinamento e base de teste.

Foram obtidos 5 conjuntos (bases) de treinamento/teste, através da seleção randômica dos vetores de características da base, na proporção 50%/50%. Os valores estimados para os parâmetros C e γ, para cada base de treinamento, utilizando a função de base radial (RBF) no classificador SVM são apresentados na Tabela 4.3.

Tabela 4.3: Parâmetros usados pelo classificador SVM, obtidos a partir das características extraídas com o Índice de Diversidade de McIntosh calculado a partir da GLCM

	Conjunto de Treinamento	C	γ
Amostras não equalizadas	1	32768,00	0,12500
	2	8192,00	0,12500
	3	32768,00	0,03125
	4	32768,00	0,12500
	5	32768,00	0,03125
Amostras equalizadas	1	32768,00	2,0
	2	32768,00	2,0
	3	32768,00	2,0
	4	8192,00	2,0
	5	32768,00	2,0

Após a estimação dos parâmetros C e γ, segue-se a etapa de classificação e validação dos resultados. A Tabela 4.4 apresenta os resultados obtidos, a partir das bases de teste, com a aplicação dos parâmetros acima no classificador SVM.

Tabela 4.4: Resultados obtidos na classificação de amostras em massa e não massa utilizando o Índice de Diversidade de McIntosh a partir da GLCM.

	Conjunto de Teste	VP	FP	VN	FN	Sensibilidade	Especificidade	Acurácia
Amostras não equalizadas	1	789	134	787	140	84,93%	85,45%	85,19%
	2	***804***	***131***	***786***	***129***	***86,17%***	***85,71%***	***85,95%***
	3	796	148	769	137	85,32%	83,86%	84,59%
	4	781	122	802	145	84,34%	86,80%	85,57%
	5	799	182	749	120	86,94%	80,45%	83,68%
Amostras equalizadas	1	822	103	823	102	88,96%	88,88%	88,92%
	2	818	96	832	104	88,72%	89,66%	89,19%
	3	841	104	816	89	90,43%	88,70%	89,57%
	4	821	95	814	120	87,25%	89,55%	88,38%
	5	***839***	***101***	***820***	***90***	***90,31%***	***89,03%***	***89,68%***

Com as amostras não equalizadas, o melhor resultado obtido, de acordo com a Tabela 4.4, alcança a acurácia de 85,95%, com sensibilidade de 86,17% e especificidade de 85,71%.

Com as amostras equalizadas, o melhor resultado obtido alcança a acurácia de 89,68%, com 90,31% de sensibilidade e 89,03% de especificidade.

A equalização de histograma, como pode ser constatado na Tabela 4.4, contribuiu para melhorar a discriminação dos tecidos, uma vez que a acurácia média de 84,99%, nas amostras não equalizadas, subiu para 89,15%, nas amostras equalizadas, o que representa uma melhoria de 4%, em média, na taxa de acerto total de classificação. Como pode ser verificado, também, na Tabela 4.4, houve uma razoável queda, nas amostras equalizadas, em relação às não equalizadas, no número de pacientes sadios diagnosticados erradamente com a presença de massas (falsos-positivos), o que reduz a quantidade de biópsias.

Esta abordagem mostra um considerável progresso em relação à abordagem anterior, pois sua melhor acurácia, de 89,68%, representa 24,60% de acertos totais a mais que a abordagem usando histograma. Por alcançar uma acurácia máxima próxima a 90%, significa que esta abordagem tende a acertar o diagnóstico de 9 a cada 10 pacientes, o que demonstra um bom desempenho para discriminação das regiões mamográficas.

Com a aplicação do método *stepwise* sobre a base de características obtidas de amostras equalizadas, foram selecionadas 18 variáveis, com o melhor resultado alcançando a acurácia de 89,35%, sensibilidade de 89,86% e especificidade de 88,83%, o que revela um desempenho idêntico ao uso de todas as variáveis.

4.1.3 Abordagem usando GLRLM

A abordagem de extração de características usando o Índice de Diversidade de McIntosh calculado a partir da GLRLM (Seção 3.1.3.3) gera um vetor composto de 24 características (6 quantizações x 4 direções).

A base de características foi organizada em dois grupos de mesmo tamanho, para utilização no classificador SVM: base de treinamento e base de teste. Foram obtidos 5 conjuntos de treinamento/teste, através da seleção randômica dos vetores de características da base. Os valores estimados para os parâmetros C e γ, para cada conjunto de treinamento, utilizando a função de base radial (RBF) no classificador SVM, são apresentados na Tabela 4.5.

Tabela 4.5: Parâmetros usados pelo classificador SVM, obtidos a partir das características extraídas com o Índice de Diversidade de McIntosh calculado a partir da GLRLM

	Conjunto de Treinamento	C	γ
Amostras não equalizadas	1	32768,00	2,0
	2	8192,00	2,0
	3	32768,00	2,0
	4	8192,00	2,0
	5	32768,00	2,0
Amostras equalizadas	1	32768,00	8,0
	2	32768,00	8,0
	3	32768,00	8,0
	4	32768,00	8,0
	5	8192,00	8,0

Após a estimação dos parâmetros C e γ, segue-se a etapa de classificação e validação dos resultados. A Tabela 4.6 apresenta os resultados obtidos pelo classificador SVM utilizando os parâmetros acima.

Tabela 4.6: Resultados obtidos na classificação de amostras em massa e não massa utilizando o Índice de Diversidade de McIntosh a partir da GLRLM.

	Conjunto de Teste	VP	FP	VN	FN	Sensibilidade	Especificidade	Acurácia
Amostras não equalizadas	1	826	83	851	90	90,17%	91,11%	90,65%
	2	836	92	850	72	92,07%	90,23%	91,14%
	3	***844***	***79***	***845***	***82***	***91,14%***	***91,45%***	***91,30%***
	4	830	95	855	70	92,22%	90,00%	91,08%
	5	838	59	845	108	88,58%	93,47%	90,97%
Amostras equalizadas	1	***794***	***198***	***737***	***121***	***86,78%***	***78,82%***	***82,76%***
	2	769	170	745	166	82,25%	81,42%	81,84%
	3	774	183	737	156	83,23%	80,11%	81,68%
	4	774	171	731	174	81,65%	81,04%	81,35%
	5	770	195	736	149	83,79%	79,05%	81,41%

Com as amostras não equalizadas, o melhor resultado obtido alcança a acurácia de 91,30%, sensibilidade de 91,14% e especificidade de 91,45%. A acurácia média dos cinco conjuntos de amostras não equalizadas é de 91,03%. Para as amostras equalizadas, o melhor resultado é de 82,76% de acurácia, 86,78% de sensibilidade e 78,82% de especificidade, sendo a acurácia média, para estas amostras, de 81,81%.

O uso da equalização de histograma, nesta abordagem, mostrou-se negativo, uma vez que acarretou uma queda média de 9,22% na taxa de acerto total de classificação, em relação às amostras não equalizadas, aumentando, consideravelmente, as taxas de falsos-positivos, o

que ocasiona um aumento na quantidade de biópsias desnecessárias, e as taxas de falsos-negativos, o que retira o tratamento de uma quantidade maior de pacientes que precisam. É provável que a equalização esteja tornando a distribuição das corridas de cinza mais homogêneas (mais próximas) entre as amostras de massa e não massa, deixando as texturas destas amostras mais parecidas, o que, para muitos casos, está acarretando índices de diversidade mais próximos, em ambas as quantizações, gerando, desta forma, muitos *outliers* para o classificador e, consequentemente, uma taxa maior de erros de classificação.

Com a aplicação do método *stepwise* sobre a base de características obtidas de amostras não equalizadas, foram selecionadas 16 variáveis, com acurácia máxima de 88,49%, sensibilidade de 89,08% e especificidade de 87,90%. Portanto, o uso de todas as variáveis é mais eficiente por apresentar taxa de acerto total superior em 2,81%.

4.1.4 Abordagem usando GLGLM

A abordagem de extração de características usando o Índice de Diversidade de McIntosh calculado a partir da GLGLM (Seção 3.1.3.4) gera um vetor composto de 24 características (6 quantizações x 4 direções).

A base de características foi dividida em dois subconjuntos, para utilização no classificador SVM: base de treinamento e base de teste. Foram obtidos 5 conjuntos (bases) de treinamento/teste, através da seleção randômica dos vetores de características da base, na proporção 50%/50%. Os valores estimados para os parâmetros C e γ, para cada base de treinamento, utilizando a função de base radial (RBF) no classificador SVM, são apresentados na Tabela 4.7.

Tabela 4.7: Parâmetros usados pelo classificador SVM, obtidos a partir das características extraídas com o Índice de Diversidade de McIntosh calculado a partir da GLGLM

	Conjunto de Treinamento	C	γ
Amostras não equalizadas	1	32768,00	0,5
	2	32768,00	0,5
	3	32768,00	0,5
	4	32768,00	0,5
	5	32768,00	0,5
Amostras equalizadas	1	32768,00	8,0
	2	32768,00	8,0
	3	32768,00	8,0
	4	32768,00	8,0
	5	32768,00	8,0

Seguimos com a etapa de classificação, na qual aplicamos os parâmetros C e γ, acima, no classificador SVM, e as métricas de validação, obtendo os resultados apresentados na Tabela 4.8.

Tabela 4.8: Resultados obtidos na classificação de amostras em massa e não massa utilizando o Índice de Diversidade de McIntosh a partir da GLGLM.

	Conjunto de Teste	VP	FP	VN	FN	Sensibilidade	Especificidade	Acurácia
Amostras não equalizadas	1	767	133	772	178	81,16%	85,30%	83,19%
	2	763	161	768	158	82,84%	82,67%	82,76%
	3	***805***	***155***	***736***	***154***	***83,94%***	***82,60%***	***83,30%***
	4	785	158	752	155	83,51%	82,64%	83,08%
	5	731	153	790	176	80,60%	83,78%	82,22%
Amostras equalizadas	1	774	152	772	152	83,59%	83,55%	83,57%
	2	763	126	797	164	82,31%	86,35%	84,32%
	3	***782***	***128***	***790***	***150***	***83,91%***	***86,06%***	***84,97%***
	4	749	168	785	148	83,50%	82,37%	82,92%
	5	794	134	771	151	84,02%	85,19%	84,59%

Com as amostras não equalizadas, o melhor resultado obtido, de acordo com a Tabela 4.8, atinge a acurácia de 83,30%, sensibilidade de 83,94% e especificidade de 82,60%. Para as amostras equalizadas, o melhor resultado é de 84,97% de acurácia, 83,91% de sensibilidade e 86,06% de especificidade.

Nesta abordagem, o emprego da equalização de histograma proporciona uma taxa de acerto total que é, em média, 1,16% superior à taxa de acerto total de classificação das amostras não equalizadas. No entanto, a melhor taxa de acerto total desta abordagem, de 84,97%, mostra-se inferior, em 6,33%, à melhor taxa de acerto total da terceira abordagem (GLRLM), que é de 91,30%.

Com a aplicação do método *stepwise* sobre a base de características obtidas de amostras equalizadas, foram selecionadas 12 variáveis, com acurácia máxima de 83,68%, sensibilidade de 80,02% e especificidade de 87,38%. Portanto, o uso de todas as variáveis é mais eficiente por apresentar maior sensibilidade e maior taxa de acerto total.

4.1.5 Abordagem usando GLRLM e GLCM

A abordagem de extração de características usando o Índice de Diversidade de McIntosh calculado a partir da GLRLM (Seção 3.1.3.3) e da GLCM (Seção 3.1.3.2), com

amostras não equalizadas, gera um vetor composto da concatenação do vetor da GLRLM (24 características) com o vetor da GLCM (120 características), totalizando 144 características.

A base de características foi organizada em dois grupos de mesmo tamanho, para utilização no classificador SVM: base de treinamento e base de teste. Foram obtidos 5 conjuntos de treinamento/teste, através da seleção randômica dos vetores de características da base. Os valores estimados para os parâmetros C e γ, para cada conjunto de treinamento, utilizando a função de base radial (RBF) no classificador SVM são apresentados na Tabela 4.9.

Tabela 4.9: Parâmetros usados pelo classificador SVM, obtidos a partir das características extraídas com o Índice de Diversidade de McIntosh calculado a partir da GLRLM e GLCM

Conjunto de Treinamento	C	γ
1	32768,00	0,125
2	32768,00	0,125
3	32768,00	0,125
4	32768,00	0,125
5	32768,00	0,125

Após a estimação dos parâmetros C e γ, segue-se a etapa de classificação e validação dos resultados. A Tabela 4.10 apresenta os resultados obtidos pelo classificador SVM utilizando os parâmetros acima.

Tabela 4.10: Resultados obtidos na classificação de amostras não equalizadas, utilizando o Índice de Diversidade de McIntosh a partir da GLRLM e GLCM.

Conjunto de Teste	VP	FP	VN	FN	Sensibilidade	Especificidade	Acurácia
1	847	69	859	75	91,87%	92,56%	92,22%
2	892	64	827	67	93,01%	92,82%	92,92%
3	837	54	881	78	91,48%	94,22%	92,86%
4	860	63	865	62	93,28%	93,21%	93,24%
5	***884***	***53***	***849***	***64***	***93,25%***	***94,12%***	***93,68%***

O melhor resultado obtido nesta abordagem, de acordo com a Tabela 4.10, atinge a acurácia de 93,68%, sensibilidade de 93,25% e especificidade de 94,12%. A acurácia média dos cinco conjuntos de teste é de 92,98%, o que representa um aumento de 1,95% na taxa média de acerto total obtida com a terceira abordagem (GLRLM), que é de 91,03%.

As taxas de falsos-positivos e falsos-negativos são consideravelmente baixas, em relação às taxas VP e VN, o que revela uma promissora capacidade da abordagem para descrever a textura dos tecidos massas e não massas.

Com a aplicação do método *stepwise* sobre a base de características, foram selecionadas 26 características, com acurácia máxima de 89,62%, sensibilidade de 88,25% e especificidade de 91,03%. Portanto, o uso de todas as variáveis é mais eficiente por apresentar taxa de acerto total superior em 4,06%.

4.1.6 Abordagem usando GLRLM e GLGLM

A abordagem de extração de características usando o Índice de Diversidade de McIntosh calculado a partir da GLRLM (Seção 3.1.3.3) e da GLGLM (Seção 3.1.3.4), com amostras não equalizadas, gera um vetor composto de 48 características, resultante da concatenação do vetor da GLRLM (24 características) com o vetor da GLGLM (24 características).

A base de características foi organizada em dois grupos de mesmo tamanho, para utilização no classificador SVM: base de treinamento e base de teste. Foram obtidos 5 conjuntos de treinamento/teste, através da seleção randômica dos vetores de características da base. Os valores estimados para os parâmetros C e γ, para cada conjunto de treinamento, utilizando a função de base radial (RBF) no classificador SVM, são apresentados na Tabela 4.11.

Tabela 4.11: Parâmetros usados pelo classificador SVM, obtidos a partir das características extraídas com o Índice de Diversidade de McIntosh calculado a partir da GLRLM e GLGLM

Conjunto de Treinamento	C	γ
1	8192,00	0,5
2	32768,00	0,125
3	32768,00	0,5
4	32768,00	0,5
5	8192,00	0,5

Seguimos com a etapa de classificação, na qual aplicamos os parâmetros C e γ, acima, no classificador SVM, e as métricas de validação, obtendo os resultados apresentados na Tabela 4.12.

Tabela 4.12: Resultados obtidos na classificação de amostras não equalizadas, utilizando o Índice de Diversidade de McIntosh a partir da GLRLM e GLGLM.

Conjunto de Teste	VP	FP	VN	FN	Sensibilidade	Especificidade	Acurácia
1	864	60	844	82	91,33%	93,36%	92,32%
2	848	58	862	82	91,18%	93,70%	92,43%
3	851	88	838	73	92,10%	90,50%	91,30%
4	847	63	864	76	91,77%	93,20%	92,49%
5	***845***	***67***	***869***	***69***	***92,45%***	***92,84%***	***92,65%***

O melhor resultado obtido, de acordo com a Tabela 4.12, alcança a acurácia de 92,65%, sensibilidade de 92,45% e especificidade de 92,84%. A acurácia média dos cinco conjuntos de teste é de 92,24%, o que representa uma queda de 0,74%, em média, de acertos totais em relação à abordagem anterior (GLRLM e GLCM), que é de 92,98%. No entanto, a acurácia desta abordagem é, em média, 1,21% melhor que a da terceira abordagem (GLRLM).

Da mesma forma que a terceira abordagem (GLRLM) e quinta abordagem (GLRLM e GLCM), as taxas de falsos-positivos e falsos-negativos desta abordagem, são baixas, revelando um excelente desempenho para discriminação dos tecidos massas e não massas.

Com a aplicação do método stepwise sobre a base de características, foram selecionadas 24 características, com o melhor resultado alcançando a acurácia de 90,00%, sensibilidade de 91,13% e especificidade de 88,90%. Portanto, o uso de todas as variáveis é mais eficiente por apresentar taxa de acerto total superior em 2,65%.

4.2 Resultados Finais

Esta seção tem por objetivo discutir os principais resultados obtidos durante os experimentos com as seis abordagens empregadas. Um resumo dos melhores resultados, de acordo com a acurácia máxima, obtidos para cada abordagem é apresentada na Tabela 4.13

Tabela 4.13: Acurácia máxima obtida em cada abordagem empregada neste trabalho.

Abordagem	Amostras	Total de Variáveis	Sensibilidade	Especificidade	Acurácia
Histograma	*Equalizadas*	6	68,77%	55,86%	62,16%
GLCM	*Equalizadas*	120	90,31%	89,03%	89,68%
GLRLM	*Não Equalizadas*	24	91,14%	91,45%	91,30%
GLGLM	*Equalizadas*	24	83,91%	86,06%	84,97%
GLRLM e GLCM	*Não Equalizadas*	144	***93,25%***	***94,12%***	***93,68%***
GLRLM e GLGLM	*Não Equalizadas*	48	92,45%	92,84%	92,65%

A primeira abordagem, onde o Índice de Diversidade de McIntosh é calculado a partir do Histograma, apresentou o menor desempenho de todas, com uma baixa capacidade para descrever a textura de tecidos massa (sensibilidade de 68,77%) e tecidos não massa (especificidade de 55,86%), e uma taxa de acerto total de 62,16%.

Seguindo na análise dos resultados finais, podemos verificar na Tabela 4.13 que a segunda abordagem, onde o índice de diversidade é calculado a partir da GLCM, e a terceira abordagem, onde o índice de diversidade é calculado a partir da GLRLM, apresentam resultados quase idênticos: sensibilidade de 90,31% e 91,14%, especificidade de 89,03% e 91,45%, e acurácia de 89,68% e 91,30%, respectivamente. No entanto, além de apresentar resultados melhores, há outras vantagens da terceira abordagem (GLRLM) em relação à segunda (GLCM), que é a quantidade reduzida de variáveis utilizada, de 24 contra 120 da GLCM, o que reduz a carga de processamento do classificador, e a capacidade de discriminação sem a necessidade de pré-processamento com equalização de histograma.

De acordo com os resultados da Tabela 4.13, o desempenho da quinta abordagem, onde o cálculo do índice de diversidade de McIntosh é feito a partir da GLRLM e GLCM, foi superior às demais abordagens apresentadas neste trabalho, atingindo a acurácia máxima de 93,68%. A sensibilidade máxima de 93,25%, nesta abordagem, mostra-se promissora, pois a taxa de pacientes doentes (com massas) diagnosticados erradamente como sadios (falsos-negativos) é de 6,75%, ou seja, de praticamente 7 pacientes a cada 100. Da mesma forma, a especificidade máxima de 94,12%, na mesma abordagem, revela uma baixa taxa de pacientes sadios (sem massas) diagnosticados erradamente com massas (falsos-positivos), de 5,88% (6 pacientes a cada 100). Estes dados revelam, portanto, a considerável capacidade desta abordagem para discriminar tecidos massa e não massa.

A sexta abordagem, que usa GLRLM e GLGLM, apresenta um desempenho idêntico ao da quinta abordagem (GLRLM e GLCM), como pode ser verificado na Tabela 4.13, onde sua taxa de acerto total, de 92,65%, é 1,03% inferior, sua sensibilidade é 0,8% inferior, e sua especificidade é 1,28% inferior, configurando-se, desta forma, como a segunda melhor abordagem. No entanto, considerando que os melhores resultados, em todas as abordagens, foram obtidos com o uso de todas as variáveis, podemos destacar a vantagem da sexta abordagem sobre a quinta, que é o uso de menos variáveis, 48 contra 144 da quinta abordagem, para discriminar as amostras.

4.2.1 Comparação com outros trabalhos relacionados

A Tabela 4.14 apresenta uma breve comparação entre os resultados encontrados neste trabalho e alguns trabalhos citados na Seção 1.1, que realizam a classificação de regiões extraídas de mamografias em massa e não massa.

Tabela 4.14: Comparação com alguns trabalhos referentes à classificação de tecidos extraídos de mamografias em massa e não massa.

Trabalhos	Base de Dados	Acurácia
(NUNES, SILVA e PAIVA, 2010)	DDSM	83,94%
(MOAYEDI, AZIMIFAR, *et al.*, 2010)	MIAS	96,60%
(SOUSA, 2011)	DDSM	99,88%
(MARTINS, JUNIOR, *et al.*, 2010)	DDSM	86,11%
Nossa Metodologia	***DDSM***	***93,68%***

A metologia proposta, como pode ser observado na Tabela 4.14, alcançou uma acurácia comparável com os melhores resultados encontrados na literatura recente para classificação de tecidos de mamografia nas classes massa e não massa.

5 CONCLUSÃO

Este trabalho apresentou a viabilidade do uso do Índice de Diversidade de McIntosh e Máquina de Vetores de Suporte para discriminação e classificação de regiões mamográficas em massa e não massa. Alcançamos uma capacidade promissora de descrição de textura de regiões extraídas de mamografias, com a adaptação dos conceitos de riqueza de espécies e abundância relativa, usados no estudo de diversidade ecológica, através do Índice de Diversidade de McIntosh, às estatísticas de informação de textura, como a Matriz de Co-ocorrência de Níveis de Cinza (GLCM), Matriz de Comprimentos de Corrida de Cinza (GLRLM) e Matriz de Comprimentos de Lacuna de Cinza (GLGLM).

Constatamos, inicialmente, que tratar a diversidade apenas das tonalidades (níveis de cinza) presentes na região de interesse não era eficaz para descrever a sua textura, pois a máxima acurácia atingida nesta abordagem foi de 60,25%. Posteriormente, com acurácias máximas de 93,68%, na quinta abordagem proposta, com uso conjunto da GLRLM e GLCM, e 92,65% na sexta abordagem, com uso conjunto da GLRLM e GLGLM, chegamos à conclusão de que a exploração das tonalidades juntamente com as relações espaciais (distância e direção) entre pixels, presentes na co-ocorrência de níveis de cinza, corridas de cinza e lacunas de cinza, permitiu ao Índice de Diversidade de McIntosh condensar mais informações de textura, aumentando significativamente a capacidade deste para discriminar tecidos massa e não massa, sendo decisivo para o sucesso da metodologia. Podemos verificar, ainda na fase de descrição de textura, que o emprego do pré-processamento, através da equalização de histograma, em algumas abordagens melhorou (primeira, segunda e quarta abordagem), e outras (terceira, quinta e sexta abordagem) prejudicou a discriminação. O uso da técnica de quantização uniforme, empregada em seis modos, de 256 a 8 níveis de cinza, conforme suspeitava-se, permitiu explorar, em todas as abordagens apresentadas, mais informações de textura, melhorando, consideravelmente, a discrimação dos tecidos massa e não massa.

Em todas as abordagens propostas, os melhores resultados foram alcançados com o uso de todas as variáveis da base de características. Embora a Máquina de Vetores de Suporte tenha sido eficaz na classificação dos tecidos, sugerimos, em trabalhos futuros, o emprego de outros classificadores, para que se possa avaliar o desempenho destes classificadores na tarefa de reconhecimento de padrões de massa e não massa, em regiões extraídas de mamografias e,

assim, tentar melhorar ainda mais os resultados atingidos neste trabalho. Uma vez que as amostras de massa, utilizadas neste trabalho, conservam o contorno, sugerimos, também, a expansão dos resultados de classificação de massa e não massa para a classificação de massas benignas e malignas, através, por exemplo, do acrescimo de informações de geometria computadas a partir dos contornos destas amostras. Sugerimos, também, em trabalhos futuros, o uso de outra base de imagens, de preferência uma base de mamografias digitais, como, por exemplo, a DMIST (*Digital Mammographic Imaging Screening Trial*). Quanto à seleção de variáveis, visando alcançar melhores resultados com uma base reduzida de características, sugerimos, por exemplo, o uso de Algorítmos Genéticos, ou o PCA (*Principal Component Analysis*).

Por fim, a metodologia descrita neste trabalho poderá integrar uma ferramenta CADx a ser aplicada em casos reais e atuais na detecção e tratamento de câncer de mama, podendo, também, ser parte de um sistema ou metodologia CAD no intuito de classificar regiões detectadas como suspeitas em massa e não massa.

REFERÊNCIAS

ACS - American Cancer Society. **Breast Cancer**, 2012. Disponivel em: <http://www.cancer.org>.

AZEVEDO, C. M.; PEIXOTO, J. E. **Falando sobre mamografia**. Rio de Janeiro: INCA, 1993.

BEBIS, G. et al. **Advances in Visual Computing**. [S.l.]: Springer, 2006. 901 p.

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

BURGES, C. J. A Tutorial on Support Vector Machines for Pattern Recognition. **Data Mining and Knowledge Discovery**, v. 2, p. 121-167, 1998.

CHANG, C. C.; LIN, C. J. **LIBSVM-A Library for Support Vector Machines**, 2010. Disponivel em: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.

CONCI, A.; AZEVEDO, E.; LETA, F. R. **Computação Gráfica:** Teoria e Prática. Rio de Janeiro: Campus, v. 2, 2008.

CRISTIANNI, N.; SHAWE-TAYLOR, J. **An Introduction to Support Vector Machines and Other Kernel-based Learning Methods**. [S.l.]: Cambridge University Press, 2000.

GALLOWAY, M. M. Texture Analysis using Gray Level Run Lengths. **Computer Graphics and Image Processing**, v. 4, p. 172-179, 1975.

GONZALEZ, R. C.; WOODS, R. E. **Processamento de Imagens Digitais**. [S.l.]: Edgard Blucher ltda, 2000. 508 p.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing**. New Jersey: Prentice Hall, 2002.

GULIATO, D.; DE OLIVEIRA, W. A. A.; TRAINA, C. A new feature descriptor derived from Hilbert space-filling curve to assist breast cancer classification. **IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS)**, 2010. 303-308.

HAIR, J. F. J. et al. **Análise Multivariada de Dados**. [S.l.]: Bookman, 2005.

HARALICK, R. M. Statistical and Structural Approaches to Texture. **Proceedings of the IEEE**, 67, 1979. 786-804.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. Textural Features for Image Classification. **IEEE Transactions on Systems, Man, and Cybernetics**, 3, 1973. 610-621.

HAWKINS, J. K. Textural Properties for Pattern Recognition. **In Picture Processing and Psychopictorics**, 1970. 347-370.

HEATH, M. et al. The Digital Database for Screening Mammography. **In: Proceedings of the Fifth International Workshop on Digital Mammography, Medical Physics Publishing**, 2000.

INCA - Instituto Nacional do Câncer. **Mama:** Detecção Precoce, 2011a. Disponivel em: <http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/mama/deteccao_precoce>.

INCA - Instituto Nacional do Câncer. **O que é câncer**, 2011b. Disponivel em: <http://www2.inca.gov.br/wps/wcm/connect/cancer/site/oquee>.

INCA - Instituto Nacional do Câncer. **Programa Nacional de Controle do Câncer de Mama**, 2011c. Disponivel em: <http://www2.inca.gov.br/wps/wcm/connect/acoes_programas/site/home/nobrasil/programa_controle_cancer_mama/>.

INCA - Instituto Nacional do Câncer. **Tipos de Câncer:** Mama, 2011d. Disponivel em: <http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/mama>.

INCA - Instituto Nacional do Câncer. **Estimativas 2012:** Incidência de Câncer no Brasil, 2012. Disponivel em: <http://www.inca.gov.br/estimativa/2012>.

JAHNE, B. **Digital Image Processing**. [S.l.]: Springer-Verlag, 2005. 607 p.

JUNIOR, G. B. **Classificação de Regiões de Mamografias em Massa e Não Massa usando Estatística Espacial e Máquina de Vetores de Suporte**. Tese de Mestrado em Engenharia Elétrica. UFMA. São Luís. 2008.

KENNEDY, A. C.; SMITH, K. L. Soil microbial diversity and the sustaunability of agricultural soils. **Plant and Soil**, 170, 1995. 75-86.

KWONG, J. N. S.; GONG, S. Learning Support Vector Machines for a Multi-View Facel Model. **In British Machine Vision Conference**, 1999. 503-512.

LLOYD, M.; INGER, R. F.; KING, F. W. On the diversity of reptile and amphibian species in a bornean rain forest. **The American Naturalist**, 102, 1968. 497-515.

LOONEY, C. G. **Pattern Recognition using Neural Networks:** Theory and Algorithms for Engineers and Scientists. New York: Oxford University Press, 1997.

LORENA, A. C.; CARVALHO, A. C. P. L. F. Uma Introdução às Support Vector Machines. **Revista de Informática Teórica e Aplicada**, v. 14, p. 43-67, 2007.

MAGURRAN, A. E. **Measuring Biological Diversity**. [S.l.]: Blackwell Science Ltd, 2004. 248 p.

MAHAFEE, W. F.; KLOEPPER, J. W. Temporal changes in the bacterial communities of soil, rhizosphere, and endorhiza associated with field-grown cucumber (Cucumis sativus L.). **Microbial Ecology**, v. 34, p. 210-223, 1997.

MAMOWEB, 2011. Disponivel em: <http://lapimo.sel.eesc.usp.br/lapimo/portal/index.html>.

MARTINS, O. L. et al. Comparison of Support Vector Machines and Bayesian Neural Networks Performance for Breast Tissues using Geostatistical Functions In Mammographic Images. **International Journal of Computational Intelligence and Applications**, 9, 2010. 271-288.

MASCARO, A. A. **Segmentação de Imagens de Mamografias Digitais**. Monografia de Graduação em Engenharia da Computação. POLI. Recife. 2007.

MATARREDONA, E. **O Uso do Processamento de Imagens Aplicadas na Radiologia**. [S.l.]: [s.n.], 1994.

MCINTOSH, R. P. An Index of Diversity and the Relation of Certain Concepts to Diversity. **Ecological Society of America**, 48, 1967. 392-404.

MELLO-THOMS, C. Interactive Computer-Aided Diagnosis of Breast Masses: Computerized Selection of Visually Similar Image Sets From a Reference Library. **Academic Radiology**, 14, 2007. 917-927.

MELO, A. S. O que ganhamos confundindo riqueza de espécies e equabilidade em um índice de diversidade? **Biota Neotrop.**, 8, 2008. 21-27.

MERT, A.; KILIC, N.; AKAN, A. Breast cancer classification by using support vector machines with reduced dimension. **Proceedings Elmar**, 2011. 37-40.

MOAYEDI, F. et al. Contourlet-based mammography mass classification using the SVM family. **Computers in Biology and Medicine**, 4, 2010. 373-383.

MOHANTY, A. K.; BEBERTA, S.; LENKA, S. K. Classifying Benign and Malignant Mass using GLCM and GLRLM based Texture Features from Mammogram. **International Journal of Engineering Research and Applications**, 1, 2011. 687-693.

MORSE, B. S. Data Structures for Image Analysis. **Brigham Young University**, 2000. Disponivel em: <http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/MORSE/data-structures.pdf>.

NOBLE, W. S. What is a Support Vector Machine? **Nature Biotechnology**, 24, 2006. 1565-1567.

NUNES, A. P.; SILVA, A. C.; PAIVA, A. C. Detection of Masses in Mammographic Images using Geometry, Simpson's Diversity Index and SVM. **International Journal of Signal and Imaging Systems Engineering**, 1, 2010. 40-51.

PAPADOPOULOS, A.; FOTIADIS, D. I.; LIKAS, A. Characterization of Clustered Microcalcifications in Digitized Mammograms using Neural Networks and Support Vector Machines. **Artificial Intelligence in Medicine**, 34, 2005. 141-150.

PEDRINI, H.; SCHWARTZ, W. R. **Análise de Imagens Digitais:** Princípios, Algorítmos e Aplicações. [S.l.]: Thomson Learning, 2008. 503 p.

PIANKA, E. R. **Evolutionary Ecology**. New York: HarperCollins, 1994.

PRATT, W. K. **Digital Image Processing**. 3. ed. [S.l.]: PIKS Inside, 2001.

SANTOS, E. M. **Teoria e Aplicação de Support Vector Machines à Aprendizagem e Reconhecimento de Objetos Baseado na Aparência**. Tese de Mestrado em Informática. UFPB. Campina Grande. 2002.

SILVA, A. C. **Melhoramento de Imagens**. Universidade Federal do Maranhão. [S.l.]. 2010.

SILVA, C. A. **Caracterização de Nódulos Pulmonares Solitários utilizando Índice de Simpson e Máquina de Vetores de Suporte**. Tese de Mestrado em Engenharia Elétrica. UFMA. São Luís. 2009.

SILVA, C. A.; CARVALHO, P. C. P.; GATTASS, M. Diagnosis of Lung Nodule using Semivariogram and Geometric Measures in Computerized Tomography Images. **Computer Methods and Programs in Biomedicine**, 79, 2005. 31-38.

SOUSA, U. S. **Classificação de Massas na Mama a partir de Imagens Mamográficas usando o Índice de Diversidade de Shannon-Wiener**. Tese de Mestrado em Engenharia Elétrica. UFMA. São Luís. 2011.

VAPNIK, V. N. **Statistical Learning Theory**. New York: Wiley, 1998. 736 p.

XINLI, W.; ALBREGTSEN, F.; FOYN, B. Texture Features from Gray Level Gap Length Matrix. **IAPR Workshop on Machine Vision Applications**, 1994. 375-378.

ZHANG, P.; KUMAR, K. Analyzing Feature Significance from Various Systems for Mass Diagnosis. **International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce**, 2006. 141-141.

APÊNDICE A – Comparando o uso de amostras de tamanhos diferentes com o uso de amostras de tamanhos iguais

Neste experimento extra (complementar), foi utilizado a mesma base DDSM (Seção 3.1.1), porém com 2492 amostras. Para a construção do descritor de textura de cada amostra, foi empregado a quinta abordagem (Seção 4.1.5), onde o Índice de Diversidade de McIntosh é calculado a partir da Matriz de Co-ocorrência de Níveis de Cinza (GLCM) e da Matriz de Comprimentos de Corrida de Cinza (GLRLM). A finalidade deste experimento é comprovar a viabilidade do uso do Índice de Diversidade de McIntosh (Equação 3) para a descrição de textura de amostras de tamanhos diferentes, de modo que estas possam ser comparadas durante o processo de classificação. Para isto, foram extraídas amostras de dimensão padronizada, de 100 x 100 pixels e 50 x 50 pixels. Assim, foram realizadas classificações sobre a base de amostras de tamanhos diferentes, a base de amostras de dimensão 100 x 100, e a base de amostras de dimensão 50 x 50, cujos os melhores resultados são apresentados na Tabela A.1.

Tabela A.1: Comparação dos melhores resultados obtidos a partir de amostras de tamanhos diferentes e tamanhos padronizados.

Dimensão das Amostras	VP	FP	VN	FN	Sensibilidade	Especificidade	Acurácia
Tamanhos diferentes	***580***	***36***	***589***	***41***	***93,40%***	***94,24%***	***93,82%***
100 x 100 pixels	536	36	596	78	87,30%	94,30%	90,95%
50 x 50 pixels	560	34	568	84	86,96%	94,35%	90,53%

De acordo com a Tabela A.1, os testes com amostras de tamanhos diferentes alcançou a acurácia de 93,82%, bem próxima às acurácias de 90,95% e 90,53%, alcançadas pelas amostras de tamanhos iguais. As especificidades foram idênticas, na casa de 94%, porém a sensibilidade foi melhor, em praticamente 6%, com o uso de amostras de tamanhos diferentes, que pode ser justificado por conta dessas amostras conservarem mais informação de textura (região interna ao contorno do tecido massa). Desta forma, como os resultados obtidos são próximos, podemos concluir que o índice de diversidade de McIntosh (Equação 3) se encaixa na teoria citada na Seção 3.1.1, como um índice relativamente independente do tamanho da amostra, de modo que, com o seu uso, amostras de tamanhos diferentes podem ser comparadas, quanto às suas diversidades de espécies.