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

ANDRÉ PEREIRA NUNES

DETECÇÃO DE MASSAS EM IMAGENS MAMOGRÁFICAS USANDO ÍNDICE DE DIVERSIDADE DE SIMPSON E MÁQUINA DE VETORES DE SUPORTE

São Luís
2009

André Pereira Nunes

DETECÇÃO DE MASSAS EM IMAGENS MAMOGRÁFICAS USANDO ÍNDICE DE DIVERSIDADE DE SIMPSON E MÁQUINA DE VETORES DE SUPORTE

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

Orientador: Prof. Dr. Aristófanes Corrêa Silva

Co-orientador: Prof. Dr. Anselmo Cardoso de Paiva

São Luís
2009

Nunes, André Pereira.

Detecção de massas em imagens mamográficas usando índice de diversidade de Simpson e máquina de vetores de suporte / André Pereira Nunes. – São Luís, 2009.

83f.

Dissertação (Mestrado). Programa de Pós-Graduação em Engenharia de Eletricidade – Universidade Federal do Maranhão – UFMA, 2009.

1. Mamografia. 2. Índice de Diversidade de Simpson. 3. Método K-Means. 4. Máquina de Vetores de Suporte.

CDU 621.386.84:618.19

DETECÇÃO DE MASSAS EM IMAGENS MAMOGRÁFICAS USANDO ÍNDICE DE DIVERSIDADE DE SIMPSON E MÁQUINA DE VETORES DE SUPORTE

André Pereira Nunes

Dissertação aprovada em 20 de fevereiro de 2009.

Prof. Aristófanes Corrêa Silva, Dr.
(Orientador)

Prof. Anselmo Cardoso de Paiva, Dr.
(Co-orientador)

Profa. Aura Conci, Dra.
(Membro da Banca Examinadora)

Prof. Alexandre César Muniz de Oliveira, Dr.
(Membro da Banca Examinadora)

À minha família e amigos.

AGRADECIMENTOS

Aos meus pais, pelo carinho e cuidados que sempre recebi em todos os momentos da minha vida.

Aos meus orientadores Aristófanes Silva e Anselmo Paiva por todo o apoio, atenção e compreensão que me dedicaram durante todo o desenvolvimento deste trabalho.

Aos colegas de curso, pela força e companheirismo demonstrados ao longo desses últimos dois anos e aos amigos Jonathan Mendes e Vânia Márcia, pela presteza com que me ajudaram sempre que precisei.

A todos que direta ou indiretamente contribuíram para a realização deste trabalho.

RESUMO

O câncer de mama é uma das maiores causas de mortalidade entre as mulheres no mundo todo. Atualmente, a análise da radiografia da mama é o recurso mais utilizado na detecção precoce desse tipo de câncer, pois possibilita a identificação de anomalias em sua fase inicial, fator fundamental para o sucesso do tratamento. A sensibilidade desse tipo de exame, no entanto, depende de diversos fatores, tais como tamanho e localização das anomalias, densidade do tecido mamário, qualidade dos recursos técnicos e habilidade do radiologista. Este trabalho apresenta uma metodologia para detecção de massas em imagens digitais de mamografias que poderá auxiliar o especialista em sua análise. O método proposto utiliza o algoritmo de agrupamento *K-Means* e a técnica de *Template Matching* para segmentar as regiões suspeitas de conterem massas. Em seguida, medidas de geometria e textura são extraídas de cada uma dessas regiões, sendo a textura descrita através do Índice de Diversidade de Simpson, uma estatística usada na Ecologia para mensurar a biodiversidade de um ecossistema. Finalmente, essas informações são submetidas a uma Máquina de Vetores de Suporte para que as regiões suspeitas sejam classificadas em massas ou não massas. A metodologia foi testada com 650 imagens mamográficas obtidas da base de dados DDSM, atingindo 83,94% de acurácia, 83,24% de sensibilidade, e 84,14% de especificidade em média.

Palavras-chave: Mamografia, Detecção Auxiliada por Computador, *K-Means*, *Template Matching*, Índice de Diversidade de Simpson, Máquina de Vetores de Suporte.

ABSTRACT

Breast cancer is one of the major causes of mortality among women throughout the world. Presently, the analysis of breast radiography is the most used method to early detection of this kind of cancer. It enables the identification of anomalies at their initial stage, which is a fundamental factor for success in the treatment. The sensitivity of this kind of exam, although, depends on several factors, such as the size and the location of the abnormalities, density of the breast tissue, quality of the technical resources and radiologist's ability. This work presents a methodology that uses the K-Means clustering algorithm and the Template Matching technique for segmentation of suspicious regions. Next, geometry and texture features are extracted from each of these regions, being the texture described by the Simpson's Diversity Index, a statistic used in Ecology to measure the biodiversity of an ecosystem. Finally, this information is submitted to a Support Vector Machine so that the suspicious regions are classified into masses and non-masses. The methodology was tested with 650 mammographic images from the DDSM database, achieving 83.94% of accuracy, 83.24% of sensibility and 84.14% of specificity in average.

Keywords: Mammography, Computer-Aided Detection, K-Means, Template Matching, Simpson’s Diversity Index, Support Vector Machine.

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE SIGLAS E ABREVIATURAS

ACS	*American Cancer Society* (Sociedade Americana de Câncer)
ADL	Análise Discriminante Linear
AG	Algoritmos Genéticos
CAD	Computer-Aided Detection (Detecção Auxiliada por Computador)
CADX	Computer-Aided Diagnosis (Diagnóstico Auxiliado por Computador)
DDSM	Digital Database for Screening Mamography (Banco de Dados Digital para Análise de Mamografia)
FDA	Food and Drugs Adminstration (Administração de Alimentos e Medicamentos)
FN	Falso Negativo
FP	Falso Positivo
INCA	Instituto Nacional do Câncer
MVS	Máquina de Vetores de Suporte
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

O controle do câncer no Brasil representa, atualmente, um dos grandes desafios à saúde pública do país. As estimativas apontam que até o final de 2009 ocorrerão mais de 460 mil novos casos da doença em todo o país. Entre a população feminina, o tipo mais incidente é o câncer de mama, com 49.400 casos previstos (INCA, 2008).

Esse tipo de câncer é especialmente temido pelas mulheres, não só pelas altas taxas de incidência e de mortalidade, mas por abranger fatores associados à perda da feminilidade nos casos em que a doença acarreta na extração parcial ou mesmo total da mama afetada.

Atualmente, o recurso mais utilizado para detecção precoce do câncer de mama é o exame de mamografia, no qual a radiografia da mama é analisada por um especialista na tentativa de identificar indícios de anormalidades no tecido mamário. Esse procedimento permite a identificação visual de anomalias em seu estágio inicial, da ordem de milímetros, um fator determinante para o sucesso do tratamento.

A sensibilidade do exame de mamografia, no entanto, depende de diversos fatores, como o tamanho e a localização da lesão, a densidade do tecido mamário e a qualidade dos recursos técnicos utilizados. Além disso, a tarefa de interpretar cuidadosamente um grande número de casos demanda tempo e um nível de atenção muito grande por parte do radiologista (ACS, 2008).

Todos esses fatores motivaram o surgimento de diversas pesquisas ao longo das últimas décadas, no sentido de desenvolver sistemas computacionais para auxiliar o especialista na tarefa de interpretação das imagens radiológicas. Esses sistemas de Detecção e Diagnóstico Auxiliado por Computador – do inglês *Computer-Aided Detection* (CAD) / *Diagnosis* (CADx) – vêm ganhando cada vez mais espaço na medicina moderna, fornecendo uma segunda opinião aos especialistas e aumentando as taxas de acerto na identificação precoce de doenças graves, como o câncer de mama (FENTON *et al.* 2007).

Este trabalho apresenta uma metodologia CAD para ajudar o especialista na tarefa de detecção de massas em imagens mamográficas. A metodologia utiliza o algoritmo de agrupamento *K-Means* e a técnica de *Template Matching* para segmentar as regiões de interesse, em seguida extrai diversas medidas geométricas dessas regiões e descreve sua textura através do Índice de Diversidade de Simpson. Finalmente, um algoritmo genético é utilizado em conjunto com o método de aprendizado supervisionado Máquina de Vetores de Suporte (MVS), em um processo integrado, para selecionar as características mais relevantes e classificar as regiões candidatas em massas e não-massas. Nesse contexto, são consideradas como massas quaisquer regiões que correspondam a uma neoplasia, seja ela de natureza maligna ou benigna.

O Índice de Diversidade de Simpson é uma medida tradicionalmente utilizada na área de Ecologia para mensurar a biodiversidade de um ecossistema e a sua utilização na área de processamento de imagens médicas representa uma importante inovação, contribuindo para o desenvolvimento dos sistemas CAD com uma maneira alternativa de caracterização da textura dos tecidos mamários.

A qualidade dos resultados obtidos a partir deste trabalho poderá tornar possível a incorporação da presente metodologia em uma ferramenta para a área médica, que possa servir como uma segunda opinião ao profissional, especialmente em casos de difícil visualização e identificação das anormalidades.

1.1 Trabalhos Relacionados

A eficiência dos sistemas ou metodologias CAD/CADx é altamente dependente do método de extração de características (ZHANG e KUMAR, 2006), da seleção das características mais adequadas para uma maior discriminação e do classificador utilizado. A literatura disponível traz trabalhos reconhecidos que tratam do mesmo problema abordado pelo método proposto, qual seja, desenvolver métodos computacionais que possam auxiliar o especialista na tarefa de análise das imagens médicas.

Em (BRAZ JÚNIOR *et al*., 2007), é apresentada uma metodologia para discriminação e classificação de regiões extraídas de mamografias em massas e não-massas através de estatísticas espaciais, como o Índice de *Moran* e o Coeficiente de *Geary*. O trabalho utiliza uma Máquina de Vetores de Suporte (MVS) para classificação das regiões, obtendo 99,64% de acurácia.

Outra metodologia que também utiliza MVS para a classificação de regiões da mama, com 89,30% de acurácia, é proposta em (MARTINS, 2007). Nela o algoritmo *Growing Neural Gas* é utilizado para segmentação dos candidatos a massa e a Função *K* de *Ripley* é utilizada para descrever a textura dos mesmos.

Em (*COSTA et al.*, 2007) é comparada a eficiência dos classificadores Máquina de Vetores de Suporte (MVS) e Análise Discriminante Linear (ADL). Os resultados da classificação alcançam 89,2% e 99,6% para ADL e MVS, respectivamente, no objetivo de classificação de regiões em massa e não-massa, comprovando a maior capacidade de generalização que o MVS é capaz de realizar sobre as amostras tratadas.

Uma nova abordagem para classificação de massas em mamografias é proposta em (MOAYEDI *et al*., 2007), onde são utilizados vetores de suporte baseados em uma rede neural *fuzzy* (SVFNN – *Support Vector Based Fuzzy Neural Network*) para desempenhar as tarefas do classificador. As características são extraídas a partir da representação das massas no domínio da freqüência utilizando os coeficientes obtidos a partir de *contourlet*.

Timp *et al*. (2007) analisam o desempenho da metodologia de classificação de massas quando as características extraídas são realizadas em mamografias consecutivamente obtidas no tempo. O principal objetivo é melhorar a descrição de massas utilizando informações presentes em mais de uma mamografia, obtidas sucessivamente.

Em (OSTA *et al.*, 2008) é investigada a utilização de *wavelets* para a extração de características dos tecidos da mama. O estudo compara o desempenho da classificação das regiões de interesse em massas e não massas através de uma MVS e de uma rede neural RBF (*Radial-Basis-Function Neural Network - RBFNN*).

Em (CAMPOS *et al*., 2007), é proposta uma metodologia para discriminação e classificação de tecidos da mama nas classes benigno, maligno e normal, usando *Independet Component Analisys* (ICA) e redes neurais. O trabalho utiliza a base de dados MIAS, obtendo uma acurácia de 97%.

Os trabalhos relacionados acima indicam serem promissoras as pesquisas de novas técnicas para extração de características de imagens radiológicas, especialmente no que diz respeito à utilização de medidas geoestatísticas como descritores de textura, assim como a utilização do método de Máquina de Vetores de Suporte como classificador por apresentar resultados superiores durante a etapa de generalização de resultados.

1.2 Organização do Trabalho

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

No Capítulo 2 é exposta a fundamentação teórica necessária ao desenvolvimento da metodologia proposta. Além dos conceitos e técnicas de processamento de imagens digitais utilizadas, são descritas as características de geometria e de textura extraídas das regiões de interesse, principalmente o Índice de Diversidade de Simpson. A técnica de seleção de características baseada em busca genética e o método de classificação denominado Máquina de Vetores de Suporte também são descritos neste capítulo.

O Capítulo 3 descreve a metodologia proposta, apresentada em quatro etapas: o pré-processamento das imagens, a segmentação das regiões de interesse, a extração de características e o processo integrado de seleção de características e classificação das regiões em massas ou não-massas.

No Capítulo 4 são apresentados e discutidos os resultados obtidos com a metodologia proposta. Finalmente, o Capítulo 5 apresenta a conclusão sobre o trabalho, mostrando a eficiência dos métodos utilizados e apresentando sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica necessária para a compreensão da metodologia utilizada para alcançar os objetivos esperados. Aborda-se o câncer de mama, o exame de mamografia, as técnicas de processamento de imagens digitais utilizadas, as características de geometria e de textura, a seleção de características baseada em busca genética, o método de classificação Máquina de Vetores de Suporte e os indicadores de desempenho utilizados para validar a metodologia.

2.1 O Câncer de Mama

O câncer é caracterizado por alterações que determinam um crescimento celular desordenado. As células afetadas multiplicam-se de maneira descontrolada, invadindo tecidos vizinhos e formando tumores, como ilustrado na Figura 1.

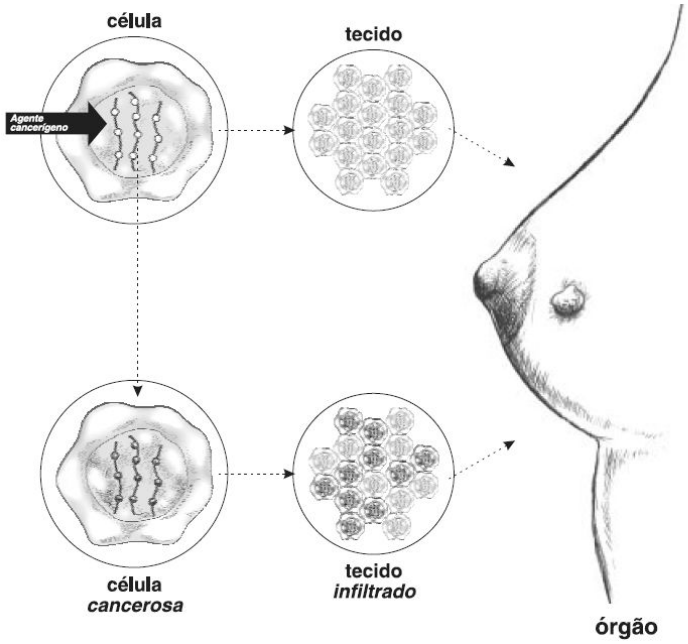

Figura 1 – O Câncer de Mama. Fonte: (MINISTÉRIO DA SAÚDE, 2002).

As células invadidas acabam perdendo a sua função especializada e, à medida que substituem as células normais, comprometem a função do órgão afetado. Podem adquirir a capacidade de se desprender do tumor e migrar, através da corrente sanguínea ou do sistema linfático, chegando a órgãos distantes, constituindo as metástases (MINISTÉRIO DA SAUDE, 2002).

O processo de formação de câncer é em geral lento, podendo levar vários anos para que uma célula prolifere e dê origem a um tumor palpável. Esse processo é composto de vários estágios, quais sejam: estágio de iniciação, onde os genes sofrem ação de fatores cancerígenos; estágio de promoção, onde os agentes oncopromotores atuam na célula já alterada; e estágio de progressão, caracterizada pela multiplicação descontrolada e irreversível da célula (MINISTÉRIO DA SAUDE, 2002).

No Brasil, as estimativas para o ano de 2009 apontam a ocorrência 466.730 casos novos de câncer. Os tipos mais incidentes, à exceção do câncer de pele do tipo não melanoma, serão os cânceres de próstata e de pulmão no sexo masculino e os cânceres de mama e de colo do útero no sexo feminino, acompanhando o mesmo perfil da magnitude observada no mundo (INCA, 2008).

O câncer de mama é provavelmente o mais temido pelas mulheres, devido à sua alta freqüência e, sobretudo, pelos seus efeitos psicológicos, que afetam a percepção da sexualidade e a própria imagem pessoal. Ele é relativamente raro antes dos 35 anos de idade, mas acima desta faixa etária sua incidência cresce rápida e progressivamente (INCA, 2008).

Epidemiologicamente ainda não foi confirmado nenhuma evidência relevante na adoção de métodos específicos de prevenção, muito embora alguns fatores ambientais e comportamentais possam estar associados a um risco aumentado de desenvolver o câncer de mama. Desta forma, a detecção precoce ainda é um instrumento imprescindível no combate à doença.

A detecção precoce se baseia na premissa de que quanto mais cedo for diagnosticado o câncer, maiores as chances de cura, a sobrevida e a qualidade de vida do paciente, além de mais favorável a relação efetividade/custo do tratamento. O objetivo é detectar as lesões pré-

cancerígenas ou mesmo o próprio câncer quando ainda localizado no órgão de origem, sem invasão de tecidos vizinhos ou outras estruturas.

A forma mais eficaz para a detecção precoce do câncer de mama é o exame radiológico da mama, a mamografia, pois permite que o especialista identifique lesões muito pequenas, em sua fase inicial, da ordem de milímetros.

2.1.1 A Mamografia

A mamografia é, atualmente, o exame com maior potencialidade comprovada para detectar o câncer de mama clinicamente oculto, em tamanho e estadiamento precoces. Por isso a radiografia da mama é o recurso mais utilizado para reduzir a mortalidade através do rastreamento de mulheres assintomáticas (KOPANS, 2000).

Grande parte da estrutura da mama é formada por tecido adiposo, que é radiolucente. Isto significa que esse tipo de tecido é permeável à incidência de raios X. Por outro lado, os principais componentes da densidade radiográfica na mamografia são os tecidos conjuntivos, que são responsáveis pela maior parte das variações grosseiras de densidade. A Figura 2 apresenta uma imagem de mamografia.

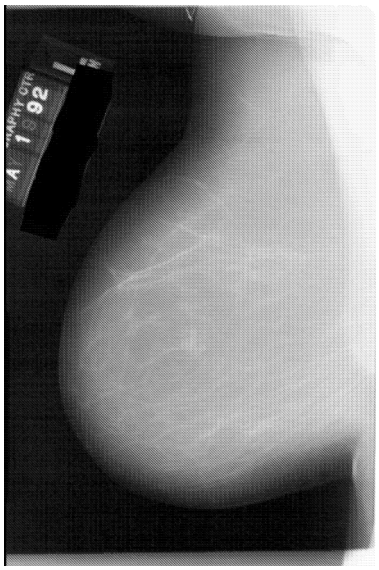

Figura 2 – Exemplo de mamografia. Fonte: (DDSM, 2001)

A mamografia deve ser feita em aparelho de raios X específico, chamado mamógrafo. Nele, a mama é comprimida de forma a fornecer melhores imagens e, portanto, melhor capacidade de diagnóstico. Para a realização do exame são necessários técnicos especializados para posicionar a paciente a fim de obter uma imagem otimizada. A compressão é necessária para evitar a subexposição da base e a superexposição dos tecidos anteriores da mama, mais finos.

Normalmente a mamografia é bilateral, ou seja, é feita uma radiografia de cada mama. Além disso, em um exame de mamografia, duas incidências ou projeções de cada mama são utilizadas: uma visão médio-lateral oblíqua e uma crânio-caudal. A Figura 3 mostra exemplos dos dois tipos de projeção.

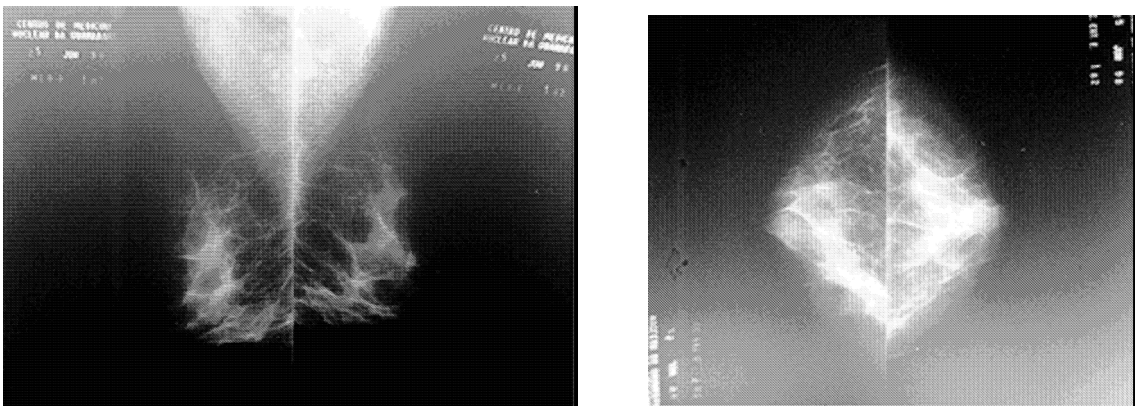

Figura 3 – (a) Mamografias com incidência Médio-Lateral (ambas as mamas); (b) Mamografias com incidência Crânio-Caudal (ambas as mamas). Fonte: (MAMOWEB, 2007).

A mamografia é um exame de alta sensibilidade. No entanto, esta sensibilidade está diretamente relacionada à idade da mulher, sendo muito menor nas mulheres jovens, que apresentam um tecido mamário bastante denso. Tipicamente, mulheres mais jovens apresentam mamas com maior quantidade de tecido glandular, o que torna esses órgãos mais densos e firmes. Ao se aproximar da menopausa, o tecido glandular vai se atrofiando e sendo substituído progressivamente por tecido gorduroso, até se constituir, quase que exclusivamente, de gordura e resquícios de tecido glandular na fase

da pós-menopausa. Essas mudanças de características promovem uma nítida diferença entre as densidades radiológicas das mamas da mulher jovem e da mulher na pós-menopausa, configurando uma dificuldade a mais para o especialista (HEATH *et al*., 1998).

Os tipos de anormalidades observáveis através da mamografia podem ser vistos na Figura 4: massas, distorções de arquitetura, e calcificações.

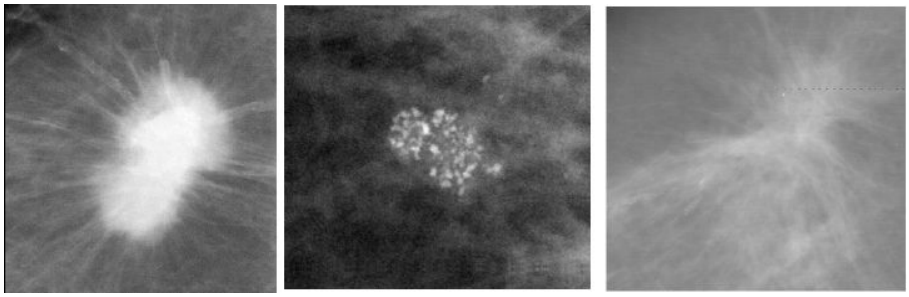

Figura 4 – Anormalidades do tecido mamário. massa espiculada (à esquerda); agrupamento de microcalcificações (no centro); distorção de arquitetura (à direita). Fonte: (DDSM, 2001).

As massas visíveis aparecem como regiões densas, de tamanho e formato variáveis. Podem ser classificadas, de acordo com o aspecto de suas bordas, como circunscritas, microlobuladas, obscurecidas, mal definidas e espiculadas, conforme mostrado na Figura 5.

Figura 5 – Classificação das massas de acordo com o aspecto de suas bordas. Adaptado de (PADWAL, 2007).

Com relação ao formato, as massas podem ser classificadas em redondas, ovais, lobulares ou irregulares, conforme mostrado na Figura 6.

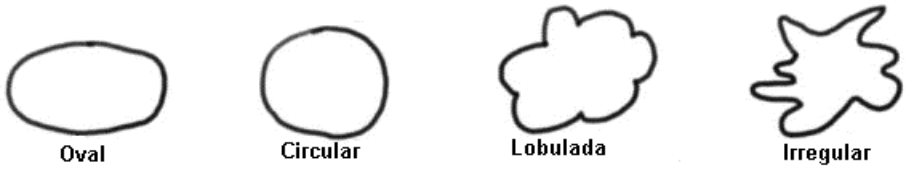

Figura 6 – Classificação das massas de acordo com sua forma. Adaptado de (PADWAL, 2007).

Embora a mamografia seja uma excelente maneira de encontrar as anormalidades da mama em sua fase inicial, momento em que as chances de cura são maiores, ela não detecta todos os casos de câncer de mama e ainda gera uma quantidade considerável de diagnósticos falso-positivos. Por isso diversas técnicas têm sido desenvolvidas para ajudar a tornar a mamografia um recurso mais preciso e eficiente. Uma das principais contribuições nessa área são os sistemas CAD/CADx, discutidos na próxima seção.

2.1.2 Sistemas de Auxílio à Detecção e Diagnóstico

Sistemas CAD/CADx (*Computer-Aided Detection/Diagnosis*), como são chamados os sistemas computacionais de auxílio à detecção e ao diagnóstico, têm a finalidade de fornecer ao especialista informações que o ajudem a tomar decisões relativas a detecção e ao diagnóstico de doenças. A diferença entre eles é que os sistemas CAD auxiliam na detecção de anormalidades sem, contudo, realizar qualquer tipo de diagnóstico sobre as mesmas, o objetivo é apenas chamar a atenção do especialista para regiões suspeitas. Por outro lado, os sistemas CADx, classificam as regiões examinadas, sugerindo o caráter benigno ou maligno das mesmas.

Em geral, os sistemas CAD/CADx fornecem opiniões a partir de informações extraídas de imagens médicas, que podem ser provenientes de diversos tipos, como a Radiografia, a Ultra-sonografia, a Ressonância Magnética, entre outras. Técnicas de Processamento de Imagens, Inteligência Artificial, Reconhecimento de Padrões, entre outras especialidades

computacionais, são aplicadas com o objetivo de melhorar tais imagens e extrair delas informações úteis à detecção de anormalidades e ao diagnóstico.

A Figura 7 apresenta as etapas principais de um sistema CAD/CADx: aquisição das regiões de interesse; caracterização das regiões de interesse por meio de textura e/ou geometria e geração do vetor de características; classificação das regiões de interesse através de algum método de reconhecimento de padrões.

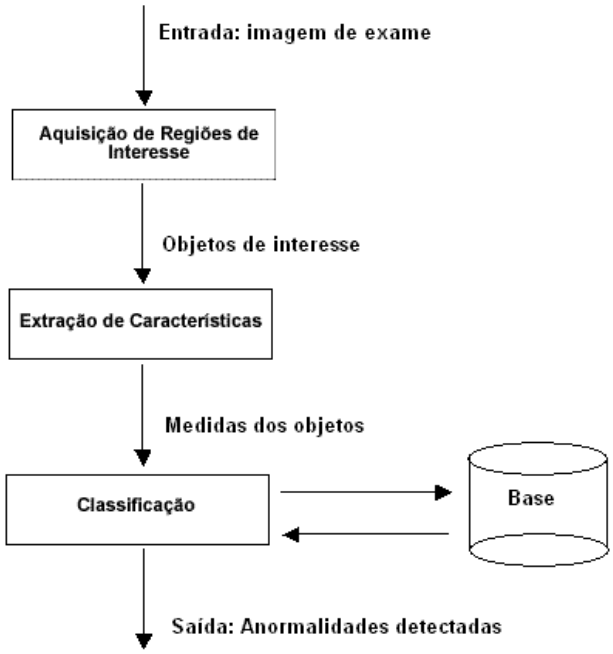

Figura 7 – Fluxo de Funcionamento de um Sistema CAD/CADx típico.

Uma vez que a anormalidade em potencial, chamada geralmente de região de interesse , é separada do restante da imagem, o sistema então extrai as principais características dessa região. Na etapa seguinte, o objeto é submetido à avaliação de um classificador, que, baseado em um treinamento realizado previamente, informa ao sistema se o objeto em questão corresponde ou não a um tecido anormal.

Diversos estudos, entre eles (FREER e ULISSEY, 2001), (BURHENNE *et al.*, 2000) e (GIGER, 2000), enfatizam a importância do uso desse tipo de sistema na detecção e diagnóstico do câncer de mama e demonstram que a

utilização dos mesmos como uma segunda opinião aumenta as taxas de detecção precoce do câncer, possibilitando aos pacientes maiores chances de cura.

Atualmente, existem sistemas CAD aprovados pela agência americana FDA (*Food and Drugs Administration*). O Programa *ImageChecker* (ROEHRIG *et al*., 1998), da empresa *R2 Technology* apresenta uma acurácia de 98,5% para detecção de calcificações e média de 0,74 falsos positivos por imagem. Para as massas, esse sistema apresenta uma sensibilidade de 85,7%, com uma taxa média de 1,32 falsos positivos por imagem. A empresa CADx *Medical Systems* desenvolveu um sistema chamado de *SecondLook*(*TM*) (iCad, 2008), o qual obteve sensibilidade de 85% para o rastreamento de cânceres (combinação de massas e microcalcificações). Adicionalmente, o sistema identificou a localização de massas malignas em 26,2% das mamografias dois anos antes do diagnóstico de câncer (SAMPAT *et al*., 2005).

A seção seguinte descreve as técnicas de processamento de imagens digitais utilizadas para desenvolver a metodologia CAD proposta neste trabalho.

2.2 Processamento de Imagens Digitais

O processamento de imagens digitais é compreendido como um conjunto de técnicas computacionais que englobam desde a aquisição da imagem digital até seu reconhecimento e interpretação por parte de uma máquina digital. Nesse contexto, uma imagem digital bidimensional é definida como uma função $f(x,y)$, que relaciona as coordenadas de um ponto (x, y) à intensidade que ele apresenta. É importante ressaltar que todos os valores de x, y e respectivas intensidades devem ser quantidades finitas e discretas, e que uma imagem digital apresenta um número finito de pontos x e y (GONZALEZ e WOODS, 2007).

O interesse em técnicas de processamento de imagens digitais surgiu, principalmente, da necessidade de melhorar a qualidade das imagens e fornecer outros subsídios que facilitem a interpretação humana. Ao longo das

duas últimas décadas, a área de processamento de imagens digitais experimentou um rápido crescimento, expandindo a cada dia o domínio de aplicações e soluções possíveis. Alguns exemplos 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.

O esquema clássico do processamento de imagens digitais é composto de várias etapas, conforme pode ser visto na Figura 8: aquisição das imagens, pré-processamento, segmentação, representação e descrição, reconhecimento e interpretação.

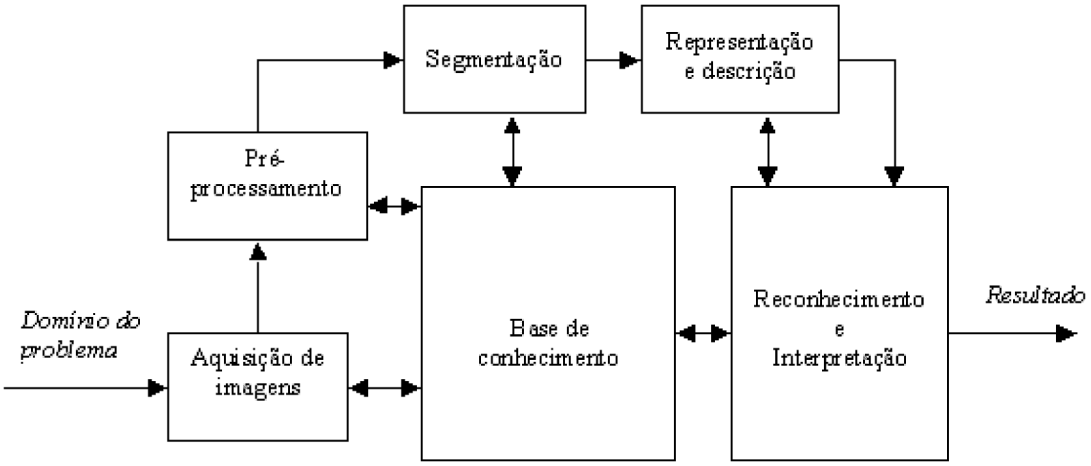

Figura 8 – Etapas fundamentais do processamento de imagens digitais. Adaptado de (GONZALEZ e WOODS, 2007).

O conjunto de resultados gerados por uma etapa é utilizado na etapa seguinte. Essas etapas formam a estrutura base para o desenvolvimento da metodologia proposta neste trabalho. A seguir é feita uma breve descrição de cada uma delas, para que se possa ter uma idéia geral do fluxo de processamento, sem se ater aos detalhes, os quais serão fornecidos no Capítulo 4.

A aquisição é o primeiro passo para o processamento de imagens digitais. Nele acontece a produção de imagens digitais de forma direta, como

em um aparelho de Raios-X digital, ou através de conversões de imagens analógicas para digitais, utilizando um digitalizador. No caso da mamografia, especificamente, os filmes impressos tradicionais podem ser digitalizados por *scanners* especializados. Para o desenvolvimento desta metodologia foi utilizada a base de mamografias digitalizadas a partir de imagens radiográficas DDSM (*Digital Database for Screening Mamography*) disponível na internet (HEATH *et al.*, 1998) (DDSM, 2001).

O passo seguinte é o de pré-processamento da imagem. O objetivo é melhorar certos aspectos da imagem, tornando mais fácil a sua identificação. Assim técnicas de realce ou melhoramento de imagens se encaixam nessa etapa. Entre elas, diminuição de ruído, realce de contraste, filtros morfológicos, dentre outras. Neste trabalho, um procedimento de remoção de ruídos baseado em algoritmos de agrupamento e de crescimento de regiões foi utilizado para remover elementos indesejáveis (identificação do paciente, fundo, etc). Além disso, um realce linear de contraste foi efetuado para aumentar a discriminação visual das estruturas da mama.

O terceiro passo na metodologia de processamento de imagem é a segmentação cujo objetivo, neste trabalho, é simplificar a imagem, reduzindo-a aos seus componentes básicos (objetos). Neste contexto, temos que segmentação é qualquer operação que faça a distinção entre os objetos contidos na imagem, ou que de alguma forma isole-os entre si. Como a segmentação é muito dependente do problema que se está abordando (tipo de imagem) não existe na literatura um método geral para se aplicar em todas as categorias de imagens. Neste trabalho, a segmentação foi feita em duas partes. Primeiro utilizando um algoritmo de agrupamento para isolar as regiões com características de intensidade semelhantes e depois selecionando somente as regiões com a forma desejada.

O quarto passo, representação e descrição, é também chamado de extração de características. Tem por objetivo determinar características básicas de cada objeto que resultem em informações importantes para discriminação entre classes distintas. O conjunto dessas medidas constitui um vetor de características que definem um padrão calculado para aquela determinada

área. Neste trabalho as regiões de interesse foram descritas através de suas características de geometria e de textura.

O objetivo da quinta etapa, reconhecimento e interpretação, é buscar, através de uma base de conhecimento (previamente construída e constituída dos padrões obtidos na etapa de representação e descrição), classificar o objeto em algum grupo determinado previamente, dependente do objetivo escolhido pelo sistema de processamento de imagem. Neste trabalho, utilizou-se uma técnica de aprendizado supervisionado para reconhecer os padrões existentes nas características de geometria e textura das regiões de interesse e classificá-las em massas ou não massas.

Nas subseções seguintes são apresentadas algumas técnicas de processamento de imagens digitais utilizadas no desenvolvimento da metodologia proposta neste trabalho.

2.2.1 Realce de Contraste

O histograma é a base para muitas técnicas de pré-processamento no domínio espacial da imagem. O histograma de uma imagem digital com intensidades variando de 0...L–1 é dado pela função discreta $f(r_k) = n_k$ onde r_k é a k-ésima intensidade e n_k é o número de *pixels* da imagem com a intensidade r_k (GONZALEZ e WOODS, 2007). A sua manipulação pode ser usada com eficiência para realizar o melhoramento de determinadas características da imagem, como o contraste.

Em uma imagem em tons de cinza, como são as imagens radiológicas, o contraste pode ser entendido como uma medida qualitativa relacionada à distribuição dos tons de cinza apresentada pela imagem. A técnica de realce de contraste tem por objetivo aumentar a discriminação visual entre os objetos presentes na imagem, sob os critérios subjetivos do olho humano. É uma técnica normalmente utilizada como uma etapa de pré-processamento para sistemas de reconhecimento de padrões.

As diversas técnicas de realce de imagens dividem-se em dois tipos de operações: pontual e local. O primeiro é caracterizado pela manipulação do

histograma da imagem, e depende somente do nível de cinza do *pixel*. Na operação local, o novo valor do *pixel* depende dos valores de seus vizinhos e inclui técnicas de filtragem, detecção de bordas e interpolação (JENSEN, 1986).

As técnicas de realce de contraste baseadas em histograma utilizam uma função matemática, chamada função de transferência, que mapeia as variações dentro do intervalo original de tons de cinza, para um outro intervalo desejado, expandindo a faixa original de valores.

Neste trabalho, foi utilizada a técnica de realce linear para aumentar a discriminação visual das estruturas presentes na mama. Ela é chamada de linear porque utiliza uma reta como função de transferência, onde apenas dois parâmetros são controlados: a inclinação da reta, que controla a quantidade de aumento de contraste, e o ponto de interseção com o eixo *X*, que controla a intensidade média da imagem final (CAMARA *et al*., 1996).

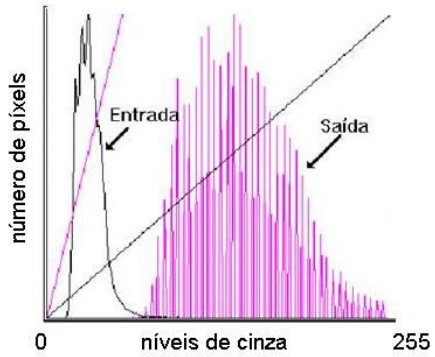

Figura 9 – Histograma original (entrada) e após o realce linear (saída).

No aumento linear de contraste as barras que formam o histograma da imagem de saída são espaçadas igualmente, uma vez que a função de transferência é uma reta. O histograma de saída será idêntico, em formato, ao histograma de entrada, exceto que ele terá um valor médio e um espalhamento diferentes, conforme mostrado na Figura 9.

2.2.2 Crescimento de Regiões

O algoritmo de crescimento de regiões é, conceitualmente, a abordagem mais simples de segmentação. O método consiste em agregar conjuntos de pixels vizinhos em regiões maiores. O processamento parte de um elemento inicial, denominado semente, o qual pode ser tanto um único pixel como um conjunto de pixels, e realiza o crescimento da vizinhança agregando os pixels próximos que possuam atributos similares aos da semente. O processo continua até que se atinja uma condição de parada pré-estabelecida, como, por exemplo, um determinado nível de cinza ou uma distância específica (PAL e PAL, 1993).

Na prática, no entanto, algumas dificuldades, razoavelmente complexas, devem ser levadas em conta durante a definição do padrão de crescimento para que resultados aceitáveis sejam obtidos, como, por exemplo, a seleção da semente, o estabelecimento das condições de semelhança e a determinação das condições de parada. Essas dificuldades, em geral, exigem que se tenha certo conhecimento *a priori* sobre a imagem que se deseja segmentar.

Neste trabalho o algoritmo de crescimento de regiões foi utilizado na etapa de pré-processamento, para remover o fundo da imagem, e na etapa de segmentação, para isolar as regiões de interesse.

2.2.3 *Template Matching*

Um dos meios mais tradicionais para a detecção de objetos em uma imagem é a técnica de *Template Matching*, na qual um modelo do objeto de interesse, chamado de *template*, é comparado com todos os objetos presentes na área da imagem. Se o *template* for suficientemente similar a um determinado objeto, diz-se que houve correspondência (*match*) entre o *template* e o objeto presente em questão (GONZALEZ e WOODS, 2007).

Essa correspondência entre o *template* e suas possíveis instâncias raramente é completamente exata por causa dos eventuais ruídos presentes na imagem e da falta de informações, *a priori*, na maioria das aplicações, sobre

a forma e estrutura dos objetos a serem detectados. Consequentemente, um procedimento comum é utilizar uma medida de similaridade entre o *template* e a região investigada. Assim uma determinada região da imagem alvo é considerada uma ocorrência do *template* sempre que a medida de similaridade calculada for maior que o limiar pré-determinado.

Neste trabalho um algoritmo de *Template Matching* foi utilizado para complementar a etapa de segmentação de regiões de interesse, descartando aquelas cujas formas não sejam minimamente parecidas com uma massa.

2.3 Características Geométricas

Durante o processamento digital de imagens, é comum extrair das regiões de interesse um conjunto de características que possam ser usadas para discriminar essas regiões adequadamente. Um tipo de característica especialmente útil para este trabalho é a informação sobre a geometria das estruturas em análise, pois a maioria das massas, mesmo as malignas, possui uma geometria bastante circular (RANGAYYAN *et al.*, 1997). Nesse sentido, quanto mais conhecidas forem as particularidades geométricas de cada objeto segmentado, maiores serão as chances de serem alcançadas taxas de classificação satisfatórias. Cinco medidas geométricas foram utilizadas neste trabalho: excentricidade, circularidade, compacidade, densidade circular e desproporção circular. As próximas subseções descrevem cada uma delas.

2.3.1 Excentricidade (E)

Excentricidade é a medida geométrica definida como a razão entre os eixos principais do objeto, caracterizando como ele está distribuído espacialmente entre seus eixos. Dessa maneira, quanto mais baixo o valor da excentricidade, mais circular é o objeto. Valores demasiadamente altos indicam uma grande diferença espacial entre o maior e o menor eixo que compõe o objeto.

A excentricidade pode ser calculada pela Equação 1:

$$E = \frac{(\mu_{02} - \mu_{20}) + 4\mu_{11}}{A} \quad (1)$$

onde *A* é a área do objeto. Os momentos centrais μ_{pq} são obtidos através da Equação 2:

$$\mu_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (x - \bar{x})^2 (y - \bar{y})^2 \quad (2)$$

com $p+q>1$ e $(\bar{x}, \bar{y})$ representando o centro de gravidade do objeto em estudo.

2.3.2 Circularidade (C)

É a medida geométrica que define o quão circular o objeto se apresenta. Trata-se da razão entre a área do objeto e o perímetro convexo, onde o perímetro convexo é calculado sobre a região que engloba o objeto de forma a não deixar picos ou defeitos no contorno. A circularidade é definida pela Equação 3:

$$C = \frac{4\pi A}{(P_{convexo})^2} \quad (3)$$

onde *A* é a área do objeto em estudo e $P_{convexo}$ é o perímetro convexo. A circularidade terá valor máximo 1 para um círculo. Logo, quanto mais próximo de um círculo é a figura, mais próximo de 1 é o valor de sua circularidade.

2.3.3 Compacidade (Co)

Essa medida mostra o quão denso o objeto é, em comparação com uma figura perfeitamente densa - o círculo (SCHOUTEN, 2003). A compacidade é definida pela Equação 4:

$$C_o = \frac{p^2}{4\pi A} \tag{4}$$

onde *A* é a área do objeto em estudo e *p* é o seu perímetro.

2.3.4 Desproporção Circular (D)

A análise desta medida, adaptada para duas dimensões por (MARTINS, 2007) a partir da medida proposta por (SOUSA *et al*., 2007) para caracterização de nódulos pulmonares em três dimensões, pode informar o quanto determinado objeto é desproporcional em relação a uma superfície totalmente circular. A desproporção circular é definida pela Equação 5:

$$D = \frac{p}{2\pi R_e} \tag{5}$$

onde *p* o perímetro do objeto em estudo e R_e é o raio estimado do círculo de mesma área do objeto em estudo. O raio estimado R_e pode ser obtido através da Equação 6:

$$R_e = \sqrt{\frac{A}{\pi}} \tag{6}$$

onde *A* é a área do objeto em estudo.

2.3.5 Densidade Circular (Dc)

Uma medida geométrica bastante comum consiste em comparar a área de um objeto com sua caixa de fronteira, isto é, o menor retângulo capaz de armazenar o objeto. Tais medidas, entretanto, enfrentam o problema de que podemos ter diferentes valores para o mesmo objeto com rotação diferente, ou seja, não são rotacionalmente invariantes.

A densidade circular, por outro lado, utiliza um círculo, figura rotacionalmente invariante, para estimar qual a porcentagem do círculo que também corresponde ao objeto (MARTINS, 2006). Isso é feito utilizando um círculo com a mesma área do objeto e centrado em seu centro de massa. A densidade circular é definida pela Equação 7:

$$D_c = \frac{100n}{A} \tag{7}$$

onde *A* é a área do objeto e *n* é o total de pontos pertencentes simultaneamente ao objeto e ao círculo de raio estimado *R*, com centro no centro de massa do objeto. A densidade circular tende a assumir valores próximos a zero para objetos muito alongados e valores próximos a 100 para objetos mais arredondados.

2.4 Características de Textura

As informações sobre a geometria dos objetos muitas vezes não são suficientes para caracterização adequada de certos tipos de objetos gráficos, pois mesmo objetos com formas muito parecidas podem ser extremamente diferentes em outros aspectos, como a textura, por exemplo.

Textura é normalmente definida como a sensação visual ou táctil que a superfície dos objetos proporciona aos sentidos. Em processamento de imagens, textura é uma informação que define a distribuição espacial de intensidades de pixels numa região da imagem (TUCERYAN e JAIN, 1998).

Uma forma clássica de quantificação da textura numa imagem em níveis de cinza é a abordagem estatística, a qual propicia a descrição da textura através das regras estatísticas que governam a distribuição e a relação entre os níveis de cinza de uma região da imagem. Medidas estatísticas comuns incluem contraste, energia, entropia, correlação, homogeneidade, momento, que são obtidas da Matriz de co-ocorrência (HARALICK *et al*., 1973).

Este trabalho propõe a utilização de uma medida de diversidade, tradicionalmente utilizada na área de Ecologia, como forma de caracterizar da textura em imagens radiológicas: o Índice de Diversidade de Simpson.

2.4.1 Índice de Diversidade de Simpson

O Índice de Diversidade de Simpson, ou simplesmente Índice de Simpson, introduzido por Edward Simpson (SIMPSON, 1949), é uma medida matemática largamente utilizada na área de Ecologia para o cálculo da biodiversidade de uma determinada comunidade, *habitat* ou região. Por biodiversidade, ou diversidade biológica, entende-se a variedade de espécies de organismos vivos que determinado ecossistema apresenta. A biodiversidade refere-se tanto ao número de diferentes categorias biológicas quanto à abundância relativa dessas categorias, e é considerada uma importante característica ecológica, uma vez que a diminuição da mesma acarreta em sérias consequências para a humanidade, como o prejuízo da atividade econômica e o próprio comprometimento funcional da natureza, na forma de alterações climáticas, deterioração do solo e das bacias hidrográficas, aumento de pragas, etc.

O cálculo do Índice de Diversidade de Simpson é feito para uma população finita de indivíduos através da Equação 8:

$$D = \frac{\sum_{i=1}^{S} n_i(n_i - 1)}{N(N-1)} \tag{8}$$

onde S representa o número total de espécies da região, N o número total de indivíduos, e n_i o número de indivíduos de uma determinada espécie i.

Na prática o Índice de Simpson mede a probabilidade de dois indivíduos selecionados aleatoriamente de uma amostra pertencerem a uma mesma espécie (ou outra categoria). O valor de D varia entre 0 e 1, onde 0 representa uma diversidade infinita e 1, ausência de diversidade. Uma forma mais intuitiva de representação, no entanto, é subtrair o valor D da unidade,

obtendo assim maiores valores de *D* quanto maior for a diversidade da região em estudo.

Uma maneira alternativa de utilização do Índice de Simpson é dividir a unidade pelo valor de *D*, obtendo assim o Índice de Reciprocidade de Simpson. Nessa abordagem o índice inicia em 1, significando uma comunidade com apenas uma espécie. Quanto maior o índice maior a diversidade, tendo como limite superior o número de espécies.

Neste trabalho, o Índice de Simpson é utilizado para medir a diversidade na distribuição dos níveis de cinza presente nas regiões de interesse de uma imagem digital de mamografia. O objetivo é utilizar o valor obtido com o índice para caracterizar a textura das regiões suspeitas, complementando as informações geométricas e fornecendo os subsídios necessários para o reconhecimento de um padrão que possibilite ao classificador distinguir as regiões que correspondem a massas e àquelas que correspondem a tecidos normais, não-massas.

2.5 Reconhecimento de Padrões

Técnicas de Reconhecimento de Padrões são usadas para classificar ou descrever padrões ou objetos através de um conjunto de propriedades ou características previamente extraídas. Um padrão é tudo aquilo para o qual existe uma entidade nomeável representante, geralmente, criada através do conhecimento cultural humano (LOONEY, 1997).

O reconhecimento de padrões envolve dois processos: classificação - onde uma amostra de uma população qualquer é particionada em grupos chamados classes; e reconhecimento - onde uma amostra desconhecida da mesma população é reconhecida como pertencente a uma das classes criadas. A classificação pode ser feita de duas formas: supervisionada e não supervisionada (LOONEY, 1997).

No processo de classificação usando aprendizagem não supervisionada, é examinado um conjunto de objetos representantes de uma população. Esse conjunto é dividido em subconjuntos (classes) de acordo com

critérios de similaridade intra-classe e dissimilaridade extra-classe. Esse processo também é chamado de agrupamento.

Por outro lado, no processo de classificação supervisionada um "reconhecedor" é treinado previamente para identificar a classe de qualquer objeto desconhecido da mesma população. Os objetos podem ser reconhecidos como pertencentes a uma determinada classe através de suas propriedades discriminantes. Um número fixo de propriedades é usado para toda população e o conjunto de seus valores determina se um objeto pertence a uma classe ou não. As propriedades individuais são chamadas de características, ou *features*, da população. Se existirem *N* características observáveis em uma população, forma-se um vetor de características. Logo os vetores de características representam os objetos em uma população de objetos. O reconhecimento de padrão é realizado através dos vetores de características.

Um passo importante é pré-processar os vetores de características a fim de retirar as características irrelevantes para a discriminação dos objetos. Se duas características são extremamente correlatas, elas são redundantes. Esse tipo de característica pode sobrecarregar o classificador e induzí-lo a erros.

Após obter as características distinguíveis de cada objeto da população, o próximo passo é atribuir um rótulo a cada um deles. O rótulo é a determinação, *a priori*, de uma classe a partir do conhecimento humano. Um conjunto de amostras, com seus rótulos e características, será usado no classificador no processo de treinamento. Nesse processo, o classificador busca gerar uma assinatura única para cada rótulo contido no conjunto de amostras. Essa assinatura será especialmente útil no processo de reconhecimento determinando o padrão identificado. Ela representa as características que melhor desempenham a distinção entre as classes.

Uma vez que o classificador esteja devidamente treinado, é possível fazer o reconhecimento do padrão de um objeto que inicialmente pertence à mesma população, mas completamente desconhecido do classificador no processo de treinamento. A técnica atribuirá um rótulo a cada objeto, a partir do conhecimento prévio obtido na etapa de treinamento, mesmo que o objeto não

pertença a nenhuma das classes. Por isso se faz necessário que a tentativa de reconhecimento de padrão de um objeto seja realizada sobre objetos da mesma população comparados aos de treinamento. Assim os padrões gerados na etapa de treinamento continuarão válidos na etapa de teste.

A metodologia proposta neste trabalho utiliza as técnicas de aprendizado supervisionado e não supervisionado em diferentes etapas do processamento. As subseções seguintes descrevem em linhas gerais as técnicas utilizadas.

2.5.1 Agrupamento de Dados (*K-Means*)

A técnica de Agrupamento de Dados consiste na classificação de objetos em diferentes grupos. Mais precisamente, é o particionamento de um conjunto de dados em subconjuntos, ou grupos, de forma que os dados dentro de cada grupo compartilhem características semelhantes em relação a alguma medida de proximidade, como a distância euclidiana, por exemplo. É uma técnica comumente utilizada para análise estatística de dados, sendo útil em diferentes áreas, incluindo aprendizado de máquinas, mineração de dados, reconhecimento de padrões, análise de imagens e bioinformática (JAIN *et al*., 1999).

O algoritmo de agrupamento utilizado neste trabalho foi o *K-Means* (HARTIGAN e WANG, 1979), um algoritmo de agrupamento não supervisionado que classifica n objetos, unicamente baseado em seus atributos, em k grupos distintos ($k<n$). Durante sua execução o *K-Means* tem por objetivo minimizar a variância total dentro dos grupos, ou seja, minimizar a função de erro quadrado, definida pela Equação 9:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left| x_j - \mu_i \right|^2 \tag{9}$$

onde S_i = 1, 2, .., k, e μ_i é a média dos elementos x_j pertencentes a S_j.

A implementação básica do *K-Means* gera aleatoriamente os *k* centros iniciais, representando os *k* grupos; a seguir atribui cada elemento ao centro mais próximo e recalcula os novos centros dos grupos de acordo com a média das distâncias entre os elementos de cada grupo. Esse processo de atribuição dos elementos e cálculo dos centros é repetido até que algum critério de parada pré-estabelecido seja atendido, como, por exemplo, a quantidade de iterações.

Neste trabalho o *K-Means* foi utilizado para segmentação das regiões de interesse, juntamente com a técnica de *Template Matching*, discutida na próxima seção.

2.5.2 Máquina de Vetores de Suporte (MVS)

A Máquina de Vetores de Suporte (MVS) – do inglês *Support Vector Machine* – introduzida por (VAPNIK, 1998), é um método de aprendizagem supervisionada usado para estimar uma função que classifique dados de entrada em duas classes. A idéia básica por trás da MVS é construir um hiperplano como superfície de decisão, de tal maneira que a margem de separação entre as classes seja máxima. O objetivo do treinamento através de MVS é a obtenção de hiperplanos que dividam as amostras de tal maneira que sejam otimizados os limites de generalização.

As MVS são consideradas sistemas de aprendizagem que utilizam um espaço de hipóteses de funções lineares em um espaço de muitas dimensões. Os algoritmos de treinamento das MVS possuem forte influência da teoria de otimização e de aprendizagem estatística. Em poucos anos, as MVS vêm demonstrando sua superioridade frente a outros classificadores em uma grande variedade de aplicações (CRISTIANINI e SHAWE-TAYLOR, 2000).

Em casos em que o conjunto de amostras é composto por duas classes separáveis, um classificador MVS é capaz de encontrar um hiperplano baseado em um conjunto de pontos, denominados *vetores de suporte*, o qual maximiza a margem de separação entre as classes. Por hiperplano entende-se uma superfície de separação de duas regiões num espaço multidimensional,

onde o número de dimensões possíveis pode ser muito grande, ou mesmo infinito. Mesmo quando as duas classes não são separáveis, a MVS é capaz de encontrar um hiperplano através do uso de conceitos pertencentes á teoria da otimização.

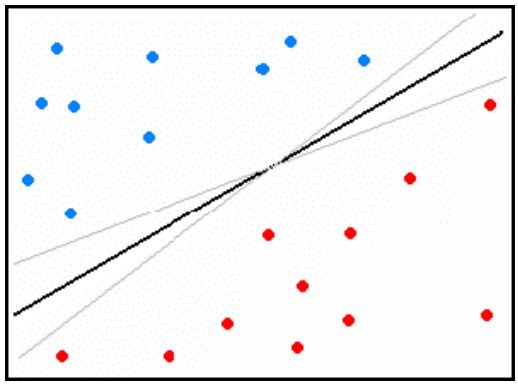

Figura 10 – Separação de duas classes através de hiperplanos.

A Figura 10 mostra em duas dimensões, para melhor visualização, hiperplanos de separação entre duas classes linearmente separáveis. O hiperplano ótimo (linha central), não somente separa as duas classes, mas mantém a maior distância possível com relação aos pontos da amostra.

Seja o conjunto de amostras de treinamento (x_i, y_i), sendo $x_i \in \Re^n$ o vetor de entrada, y_i a classificação correta das amostras e i = 1,...,n o índice de cada ponto amostral. O objetivo da classificação é estimar a função $f : \Re^n \rightarrow \{-1,1\}$, que separe corretamente os exemplos de teste em classes distintas. A etapa de treinamento estima a função $f(x) = (w.x) + b$, procurando por valores de w e b tais que a Equação 10 seja satisfeita:

$$y_i((w.x_i) + b) \geq 1 \qquad (10)$$

sendo w o vetor normal ao hiperplano de decisão e b o corte ou distância da função f em relação á origem. Os valores ótimos de w e b serão encontrados ao minimizar a Equação 11, de acordo com a restrição dada pela Equação 10 (CHAVES, 2006).

$$\Phi(w) = \frac{w^2}{2} \tag{11}$$

A MVS ainda possibilita encontrar um hiperplano que minimize a ocorrência de erros de classificação nos casos em que uma perfeita separação entre as duas classes não for possível. Isso graças a inclusão de variáveis de folga, que permitem que as restrições presentes na Equação 10 sejam quebradas. O problema de otimização passa a ser então a minimização da Equação 12, de acordo com a restrição imposta pela Equação 10, onde *C* é um parâmetro de treinamento que estabelece um equilíbrio entre a complexidade do modelo e o erro de treinamento, devendo ser selecionado pelo usuário.

$$\Phi(w,\xi) = \frac{w^2}{2} + C\sum_{i=1}^{N} \xi_i \tag{12}$$

$$y_i((w.x_i) + b) + \xi \geq 1 \tag{13}$$

Através da teoria dos multiplicadores de *Lagrange*, chega-se à Equação 14. O objetivo então passa a ser encontrar os multiplicadores de *Lagrange* α_i ótimos que satisfaçam a Equação 15 (CHAVES, 2006):

$$w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i, x_j) \tag{14}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \ , \ 0 \leq \alpha_i \leq C \tag{15}$$

Apenas os pontos onde a restrição da Equação 10 seja exatamente igual à unidade têm correspondentes $\alpha \neq 0$. Esses pontos são chamados de vetores de suporte, pois se localizam geometricamente sobre as margens. Tais pontos têm fundamental importância na definição do hiperplano ótimo, pois os mesmos delimitam a margem do conjunto de treinamento.

A Figura 11 destaca os pontos que representam os vetores de suporte. Os pontos além da margem não influenciam decisivamente na determinação do

hiperplano, enquanto que os vetores de suporte, por terem pesos não nulos, são decisivos.

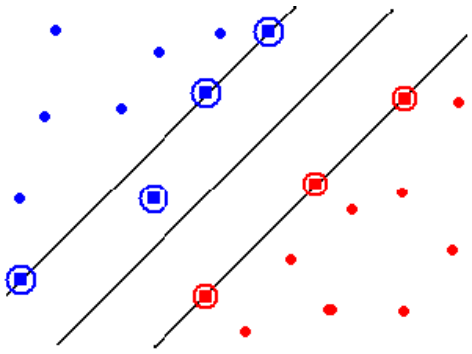

Figura 11 – Separação de duas classes através de hiperplanos.

Para que a MVS possa classificar amostras que não são linearmente separáveis, é necessária uma transformação não-linear que transforme o espaço entrada (dados) para um novo espaço (espaço de características). Esse espaço deve apresentar dimensão suficientemente grande, e através dele, a amostra pode ser linearmente separável. Dessa maneira, o hiperplano de separação é definido como uma função linear de vetores retirados do espaço de características ao invés do espaço de entrada original. Essa construção depende do cálculo de uma função *K* de núcleo de um produto interno (HAYKIN e ENGEL, 2001). A função *K* pode realizar o mapeamento das amostras para um espaço de dimensão muito elevada sem aumentar a complexidade dos cálculos. A Equação 16 mostra o resultado da Equação 14 com a utilização de um núcleo *K*.

$$w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \qquad (16)$$

Uma importante função de núcleo é a função de base radial, muito utilizada em problemas de reconhecimento de padrões e também utilizada neste trabalho. A função de base radial é definida pela Equação 17:

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \qquad (17)$$

2.6 Seleção de Características Usando Algoritmos Genéticos

Conforme dito na seção anterior, as técnicas de reconhecimento de padrões fazem uso de vetores de características extraídas dos objetos sobre os quais se deseja aprender. Teoricamente quanto mais características forem utilizadas para representar os exemplos de treinamento, mais informação estará disponível para o algoritmo de aprendizado e, portanto, melhor será o desempenho do classificador (KOLLER e SAHAMI, 1996).

Entretanto, restrições práticas sugerem que trabalhar com um conjunto reduzido de características pode ser benéfico em diversos aspectos. Em aplicações reais, por exemplo, geralmente observa-se que quanto maior o número de características, maior a quantidade de exemplos necessários para construir um classificador de bom desempenho (JAIN *et al.*, 2000). Outro problema, é que, em alguns domínios, é comum que a maioria das características disponíveis não seja informativa o suficiente para a distinção entre as diferentes classes (XING, 2003). Isto ocorre principalmente por serem irrelevantes ou redundantes ou apresentarem níveis elevados de ruído. Por isso, é comum em problemas de reconhecimento de padrões realizar uma seleção das características mais relevantes, a fim de aumentar a eficiência do classificador e diminuir os custos de processamento.

O princípio básico dos métodos de Seleção de Características (SC) consiste em, dado um conjunto de n características, selecionar um subconjunto de tamanho k ($k<n$) tal que o erro de classificação tenha a maior redução possível em relação ao conjunto completo (JAIN *et al.*, 2000).

Neste trabalho, a seleção de características foi efetuada com ajuda da técnica de Algoritmos Genéticos (AG), uma técnica de busca e otimização inspirada no processo de evolução dos seres vivos (GOLDBERG, 1989). Um AG utiliza uma população de indivíduos para resolver um dado problema. Cada indivíduo, ou cromossomo, corresponde a uma possível solução codificada do problema. Mecanismos baseados em reprodução, hereditariedade e mutação são aplicados à população atual para gerar uma nova. As populações vão evoluindo por diversas gerações até que um dado critério de parada seja atendido.

O algoritmo começa com uma população de indivíduos gerados aleatoriamente, que podem ser entendidos como tentativas iniciais de soluções para o problema. A população é avaliada e, para cada cromossomo, uma pontuação, chamada de aptidão, é dada, refletindo a qualidade da solução associada e ele. Os indivíduos mais aptos têm maiores chances de sobreviver e, assim, transmitirem seus genes para a geração seguinte.

Os conceitos fundamentais de um AG são: representação, inicialização, reprodução, seleção, mutação e substituição. A seguir uma breve descrição desses conceitos é apresentada:

- **Representação**. Os cromossomos são geralmente codificados na representação binária. Cada bit do cromossomo corresponde a um gene, que pode ter o valor 0 ou 1, indicando a ausência ou a presença da informação referente àquele gene.
- **Inicialização**. É a maneira pela qual a população inicial de cromossomos é gerada. Neste trabalho, os genes dos cromossomos foram inicializados aleatoriamente com 0 ou 1, pois não se tinha informação, *a priori*, de quais características seriam mais relevantes.
- **Reprodução**. O operador que realiza a operação de reprodução é chamado de cruzamento. Ele é usado para guiar o processo evolucionário por soluções potencialmente melhores. Ele funciona promovendo a transmissão de genes de dois indivíduos, chamados pais, para a composição de novos indivíduos, chamados filhos, tal que estes herdem as propriedades daqueles.
- **Seleção.** É o processo de escolha dos indivíduos que irão se reproduzir. Geralmente se prioriza os melhores indivíduos, baseado em suas medidas de aptidão.
- **Mutação**. É o operador genético responsável por manter a diversidade da população. Ele opera invertendo aleatoriamente algumas posições do cromossomo de acordo com alguma probabilidade pré-estabelecida. Um valor de probabilidade comumente utilizado é $1/m$, onde m é o tamanho do cromossomo.

- **Substituição**. Os esquemas de substituição determinam como as novas gerações são formadas. Neste trabalho, utilizou-se o conceito de *elitismo*, onde os *n* melhores indivíduos são perpetuados para a geração seguinte, substituindo os *n* piores. Os novos filhos gerados pelo operador de reprodução completam a nova população.

Os algoritmos genéticos têm sido largamente utilizados no problema de seleção de características devido a sua capacidade de realizar uma busca eficiente em um grande espaço de possibilidades. Neste trabalho, é utilizado um modelo de seleção de características usando um algoritmo genético integrado a um classificador MVS e ao processo de treinamento e classificação. Mais detalhes sobre a etapa de seleção de características e classificação das regiões de interesse em massas e não massas são descritos na Seção 3.6.

2.7 Medidas de Desempenho

Fazer o reconhecimento de padrões é um processo imperfeito que resulta mais em probabilidade de se estar certo do que em certeza. Existem diversas medidas para verificar o desempenho de um classificador qualquer, sendo que a medida mais importante é o desempenho do mesmo a partir da classificação de novos casos (conjunto de testes). O desempenho do classificador, medido através do conjunto de teste, é uma boa indicação de seu desempenho real.

Em problemas de processamento de imagens e reconhecimento de padrões ligados à área médica costuma-se medir o desempenho da metodologia calculando-se algumas estatísticas sobre os resultados dos testes.

Dada uma amostra com casos positivos e negativos de uma determinada doença, os resultados dos testes de classificação dos casos analisados podem ser divididos em quatro grupos: VP (Verdadeiros Positivos): número de casos corretamente classificados como positivos; FP (Falsos

Positivos): número de casos erroneamente classificados como positivos; VN (Verdadeiros Negativos): número de casos corretamente classificados como negativos; e FN (Falsos Negativos): número de casos erroneamente classificados como negativos. Esses números são utilizados para gerar medidas capazes de quantificar o desempenho de uma metodologia, para que se possa avaliar o quão eficiente ela é em atingir seus objetivos.

As medidas de desempenho mais utilizadas na área processamento de imagens médicas são: acurácia (*A*), sensibilidade (*S*), especificidade (*E*), *F-Measure* (F_m), média de falsos positivos por imagem (*FP/i*) e média de falsos negativos por imagem (*FN/i*). A acurácia mede a porcentagem total de casos corretamente classificados (Equação 18). A sensibilidade mede o desempenho da classificação em relação aos casos positivos (Equação 19). A especificidade mede o desempenho da classificação em relação aos casos negativos (Equação 20). *F-Measure* é uma medida que calcula o equilíbrio entre a sensibilidade e a especificidade da classificação, privilegiando aquelas que apresentem um bom balanceamento entre os casos de FP e FN (Equação 21). A média de falsos positivos por imagem é simplesmente a razão entre o número de falsos positivos encontrados e o total de casos avaliados, sendo a média de falsos negativos por imagem obtida de forma similar (BUSHBERG *et al.*, 2002).

$$A = \frac{VP + VN}{VP + VN + FP + FN} \quad (18) \qquad S = \frac{VP}{VP + FN} \quad (19)$$

$$E = \frac{VN}{VN + FP} \quad (20) \qquad F_m = \frac{2 \times (S \times E)}{(S + E)} \quad (21)$$

Essas seis medidas de desempenho foram utilizadas neste trabalho para validar a eficiência da metodologia em classificar as regiões suspeitas da mamografia em massas e não massas.

3 METODOLOGIA

Este capítulo descreve os procedimentos realizados pela metodologia proposta neste trabalho para detecção de massas em imagens digitais de mamografia. De forma análoga ao esquema tradicional de processamento de imagens, apresentado na Seção 2.2, a metodologia proposta também é composta de etapas, são elas: pré-processamento, segmentação das regiões de interesse, extração de características, seleção de características e classificação das regiões de interesse em massa ou não massa. A Figura 12 apresenta um diagrama ilustrando essas etapas.

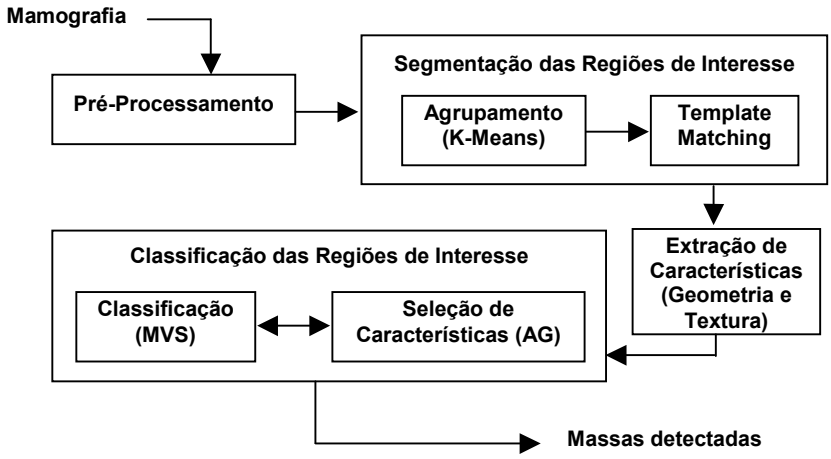

Figura 12 – Etapas da Metodologia proposta.

A etapa de pré-processamento tem o objetivo de facilitar o processamento a ser realizado pelas etapas seguintes através da remoção de elementos indesejáveis e melhoria da imagem. A etapa de segmentação das regiões de interesse identifica as áreas da imagem que são suspeitas de conterem anormalidades, de forma que as etapas seguintes trabalhem apenas com as regiões relevantes para o problema. A etapa de extração de características descreve as regiões de interesse através de suas características de geometria e textura. Finalmente, a etapa de classificação seleciona as características que melhor descrevem as regiões de interesse para treinar um classificador que seja capaz de classificar essas regiões em massas ou não massas.

O restante do capítulo descreve cada uma dessas etapas em detalhes, mas antes aborda a base de dados utilizada nos testes e os recursos de *softwares* e *hardware* utilizados para o desenvolvimento da metodologia.

3.1 Amostra

Este trabalho utilizou amostra de mamografias digitalizadas obtidas através da base de imagens pública DDSM (*Digital Database for Screening Mamography*), disponível gratuitamente na internet (HEATH *et al*., 1998). Essa base contém 2620 casos adquiridos através das seguintes instituições americanas: *Massachusetts General Hospital*, *Wake Forest University*, e *Washington University in St. Louis School of Medicine*. Cada caso contém duas imagens de cada mama, nas projeções médio-lateral oblíqua e crânio-caudal. Além disso, são disponibilizadas informações sobre a paciente, tais como a idade e a densidade da mama. Também são informados o tipo de *scanner* utilizado na digitalização do exame e a resolução da imagem. As imagens com regiões suspeitas possuem a descrição da anormalidade, além do diagnóstico e a localização da mesma. Todas as informações contidas no DDSM foram fornecidas por especialistas no assunto (HEATH *et al*., 1998).

Neste trabalho, foram utilizadas 650 imagens de mamografias, cada uma contendo apenas uma massa, escolhidas aleatoriamente da base de imagens do DDSM.

3.2 Software e Hardware Utilizados

Para a implementação dos métodos descritos neste trabalho utilizou-se a linguagem de programação *C++,* através do ambiente de desenvolvimento *Microsoft Visual C++ 2005 – Express Edition*. A manipulação das imagens foi realizada através da biblioteca de processamento de imagens *OpenCV* (INTEL, 2008) e o classificador MVS foi obtido através da biblioteca LIBSVM (CHANG e LIN, 2003), ambas disponíveis gratuitamente na internet.

O computador utilizado durante o desenvolvimento e teste da metodologia foi um *Intel Pentium D* de 2,80 GHz, com 2 GB de RAM e HD de 80 GB, rodando sistema operacional *Windows XP*.

3.3 Pré-Processamento

Muitas das imagens contidas no DDSM apresentam ruídos e outros elementos que podem interferir no processamento que se deseja realizar. Esses elementos indesejáveis incluem marcas de identificação do paciente ou do tipo de exame, *pixels* do fundo da imagem, e eventuais ruídos produzidos por imperfeições do processo de geração da imagem ou da digitalização. O objetivo desta etapa de pré-processamento é remover esses elementos indesejáveis e melhorar a discriminação visual das estruturas presentes na mama. A Figura 13 apresenta os elementos tipicamente presentes em uma imagem de mamografia digital.

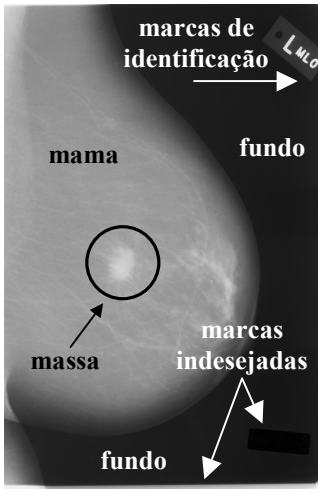

Figura 13 – Elementos presentes em uma imagem de mamografia.

O procedimento para remoção dos elementos indesejáveis utiliza o algoritmo de agrupamento *K-Means* e o algoritmo de crescimento de regiões, da seguinte forma.

Primeiro a imagem é submetida ao *K-Means* para agrupamento dos *pixels* em dois grupos (*k*=2) de acordo com suas intensidades. Isso faz com que os *pixels* de maior intensidade, como são os *pixels* da mama e os das marcas de identificação, sejam agrupados em um grupo e os de menor intensidade, como são os do fundo da imagem e dos ruídos mais escuros, sejam agrupados em outro grupo. Como o objetivo é manter apenas os *pixels* da mama, o grupo contendo os *pixels* menos intensos é eliminado, substituindo seus valores de intensidade por zero. A representação visual dos grupos gerados pode ser vista no exemplo da Figura 14.

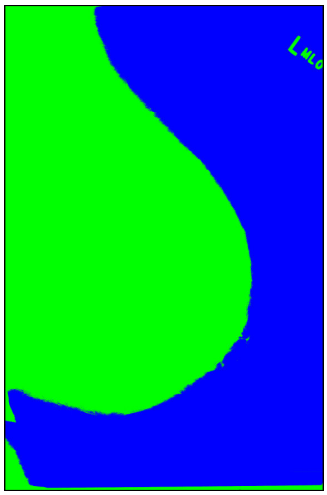

Figura 14 – Representação visual do agrupamento realizado com o *K-Means* (*k*=2). As cores representam os grupos produzidos.

O segundo passo é identificar na imagem resultante qual dos objetos presentes corresponde à mama. Para isso, é produzida uma imagem binária do grupo resultante do passo anterior, então um algoritmo de crescimento de regiões isola os objetos desconexos em imagens separadas. A partir daí seleciona-se o objeto que apresenta a maior área e descartam-se os demais. Por fim, os *pixels* originais em níveis de cinza são restabelecidos sobre a imagem resultante.

Após a remoção dos elementos indesejados, a imagem ainda passa por um processo de realce de contraste. Esse procedimento tem o objetivo de

aumentar a discriminação visual entre as estruturas presentes na mama. Neste trabalho utilizou-se o realce linear de contraste (ver Seção 2.3).

Para finalizar esta etapa a imagem é reduzida, de forma automática, a 1/3 do seu tamanho original, além de ter uma faixa de 50 *pixels* de borda removida. A redução foi feita para reduzir os custos de processamento das etapas posteriores, e a retirada da borda foi feita para minimizar os efeitos das áreas claras comumente encontradas nas laterais e nas partes superior e inferior das imagens do DDSM. A Figura 15 apresenta um exemplo de imagem resultante da etapa de pré-processamento.

Figura 15 – Imagem resultante da etapa de pré-processamento.

3.4 Segmentação das Regiões de Interesse

Esta etapa, composta de duas partes, tem o objetivo de identificar as regiões da mama com mais possibilidades de conterem massas. Na primeira parte o algoritmo de agrupamento *K-Means* é usado para agrupar os *pixels* da mama em diversos grupos, baseado em seus valores de intensidade, conforme mostrado na Figura 16 (nessa imagem, uma cor arbitrária foi atribuída a cada grupo apenas para visualização do resultado do processo de agrupamento). O objetivo é manter juntos *pixels* com características de intensidade semelhantes para, na segunda parte desta etapa, identificar aquelas estruturas agrupadas que possuam formas parecidas com massas.

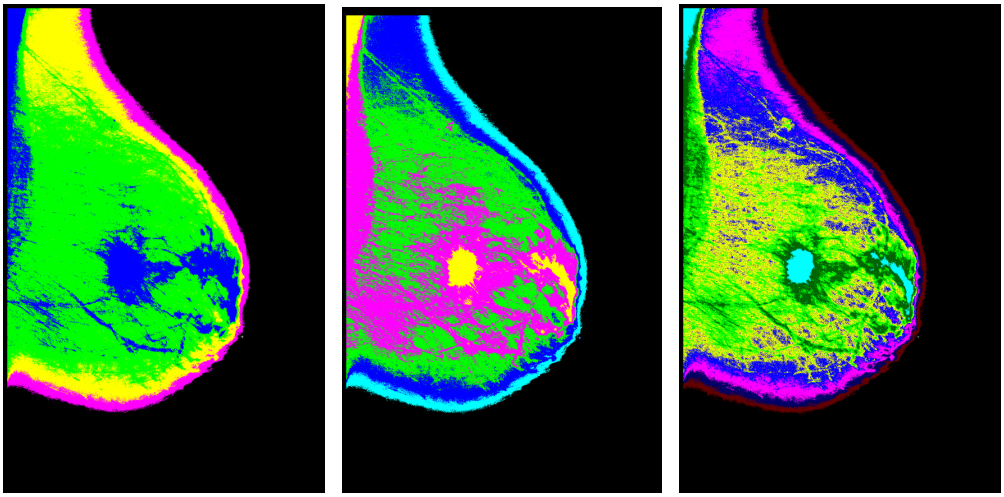

Figura 16 – Agrupamentos produzidos pelo *K-Means* para vários valores de *k*. Da esquerda para a direita: *k*=5, *k*=6, *k*=10.

Como o número ideal de grupos (*k*) que representa a distribuição natural dos *pixels* em cada imagem é, *a priori*, desconhecido, diversos valores de *k* foram utilizados (*k*=5,6,...,10). Assim, para cada imagem processada são produzidos vários grupos para cada valor de *k* utilizado. As estruturas resultantes em cada grupo são então separadas em imagens binárias individuais, usando um algoritmo de crescimento de regiões. Alguns exemplos de estruturas isoladas podem ser vistos na Figura 17.

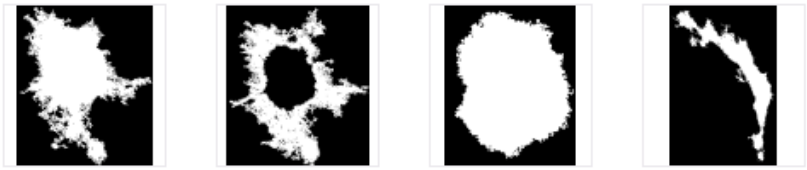

Figura 17 – Exemplos de estruturas obtidas pelo algoritmo de crescimento de regiões a partir dos grupos gerados pelo *K-Means*.

Para reduzir o grande número de estruturas segmentadas, as imagens isoladas que apresentarem dimensões inferiores a 30x30 *pixels* e superiores a 350x350 *pixels*, são descartadas. Essa faixa de valores foi determinada a partir

das informações contidas no DDSM, as quais demonstraram não haver massas fora desse intervalo na amostra selecionada.

A segunda parte da etapa de segmentação opera sobre as imagens binárias resultantes do procedimento de agrupamento e funciona como um filtro de forma. A operação consiste em selecionar apenas as estruturas minimamente parecidas com uma massa. Para isso, a técnica de *Template Matching* foi utilizada da seguinte forma. Cada estrutura isolada é percorrida por três *templates* circulares binários, de tamanhos diferentes, sendo utilizado apenas um de cada vez. Em cada posição sobreposta confere-se a quantidade de pixels que apresentam o mesmo valor do seu correspondente no *template*. Se o percentual de *pixels* coincidentes na região for superior a 70%, os *pixels* brancos da imagem alvo que coincidiram com o *template* são salvos em uma imagem resultante.

O objetivo é descartar as estruturas ocas, alongadas ou muito retorcidas. Entre as imagens da Figura 17, por exemplo, apenas as estruturas “a” e “c” são selecionadas para a próxima etapa, descartando-se as demais. A utilização de três *templates* de tamanhos diferentes foi realizada para adequar a técnica à escala das massas contidas na amostra. Isso porque nesta implementação cada *template* só consegue identificar satisfatoriamente objetos um pouco maiores ou um pouco menores que ele. Testes preliminares ajudaram a identificar os tamanhos mais adequados: 48x48, 95x95 e 158x158 *pixels*. A Figura 18 mostra os três *templates* utilizados.

Figura 18 – *Templates* binários em forma circular.

Uma vez identificadas as estruturas com forma e tamanho desejados, tem-se uma coleção de imagens isoladas de regiões suspeitas da mamografia.

No entanto, como se utiliza seis valores de *k* (*k*=6,7,8,9,10) e, em seguida, para cada agrupamento gerado três tamanhos de *template*, muitas das estruturas isoladas contém outras ou correspondem exatamente à mesma região da imagem. Para reduzir essa redundância um procedimento de filtragem foi realizado. Primeiro é criada uma imagem união, contendo todas as estruturas isoladas da mesma imagem, ainda binária, em seguida separa-se novamente essas estruturas com o mesmo algoritmo de crescimento de regiões usado para isolar as estruturas dos grupos anteriormente. Com isso as regiões sobrepostas são agrupadas em uma única região, reduzindo a redundância. Um efeito colateral indesejável dessa operação, entretanto, é o aparecimento de estruturas maiores que a faixa de valores desejada (30x30 à 350x350), o que foi resolvido aplicando-se o filtro de tamanhos novamente (30x30 à 350x350). As regiões resultantes têm então seus *pixels* originais, em níveis de cinza, restaurados, para que a etapa seguinte - extração de características - possa utilizar essas informações para descrever suas texturas.

3.5 Extração de Características

Esta etapa da metodologia tem o objetivo extrair medidas descritivas das regiões de interesse segmentadas para formar vetores de características que as representem na etapa de classificação. Para isso características de geometria e textura foram utilizadas.

A geometria das regiões de interesse é descrita através das cinco características, definidas na Seção 2.3: excentricidade, circularidade, compacidade, desproporção circular e densidade circular. O procedimento de extração dessas medidas é direto e não leva em conta as intensidades dos *pixels* das regiões de interesse, ou seja, cada *pixel* não-nulo influi da mesma maneira no cálculo de áreas, distâncias, raios, e outras medidas necessárias ao cálculo das características geométricas.

A textura das regiões de interesse é descrita através do Índice de Diversidade de Simpson (*D*). Para aplicar o conceito de diversidade a uma região da imagem, a Equação 9 foi utilizada da seguinte maneira: a variável *S*, que, na abordagem tradicional do índice, representa a quantidade total de

espécies presentes no ecossistema em estudo, passa a representar a quantidade de níveis de cinza diferentes na imagem (região de interesse); de forma análoga, *N*, passa a representar a quantidade total de *pixels*, e n_i o número de *pixels* que apresenta a intensidade *i*, o que pode ser obtido diretamente do histograma da imagem.

Para avaliar a melhor maneira de utilizar a informação de diversidade para descrever a textura das regiões de interesse, três abordagens de extração foram utilizadas: a extração global, a extração por anéis e a extração em círculos.

Na abordagem de extração global todos os *pixels* da região de interesse são levados em consideração de uma só vez para o cálculo da diversidade. Isso quer dizer que o valor *D* obtido representa a diversidade da região de interesse como um todo, independente das variações locais de diversidade que possam existir entre as diversas áreas da região.

As outras duas abordagens, pelo contrário, extraem os valores locais de diversidade nas diferentes áreas da região de interesse, na tentativa de descobrir variações nos padrões de diversidade entre as áreas mais próximas à borda da região examinada e as áreas mais internas. A diferença entre essas duas abordagens consiste no formato das áreas sobre as quais se quer medir a diversidade. Na abordagem circular, os valores de diversidade da região de interesse são extraídos a partir de *n* círculos concêntricos e sobrepostos, de diferentes raios, partindo do centro de massa da imagem. O círculo com maior raio circunscreve todos os *pixels* da região de interesse, o que equivale à mesma área coberta pela abordagem global.

O tamanho de cada raio foi definido pela expressão $R_i = i \times (R_n/n)$, com i=1,2,...,n, onde R_i é o tamanho do raio *i, n* é o número de áreas em que a região de interesse é representada, e R_n é o tamanho do maior raio, ou seja aquele que produz um círculo que circunscreve todos os *pixels* da região de interesse.

No exemplo da Figura 19, a região de interesse tem seus *pixels* tomados em três áreas circulares (n=3), de forma que um valor do Índice de Diversidade de Simpson é calculado para cada uma delas.

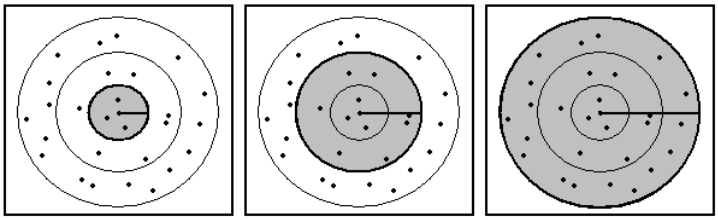

Figura 19 – *Pixels* da região de interesse tomados em áreas circulares (*n*=*3*).

A abordagem em anéis é similar à abordagem circular, mas utiliza dois raios consecutivos, em vez de um só, levando em consideração apenas os *pixels* dentro do anel formado pelos raios. No exemplo da Figura 20, a região de interesse tem seus *pixels* tomados em três áreas em forma de anéis (*n*=3), de forma que um valor do Índice de Diversidade de Simpson é calculado para cada uma delas.

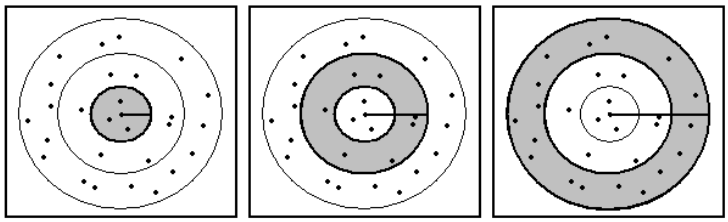

Figura 20 – Pixels da região de interesse tomados em anéis (*n*=*3*).

Obviamente, um problema dessas duas abordagens é definir o número de áreas em que a região de interesse terá seus valores de diversidade calculados. Por isso, os testes envolvendo essa abordagem utilizam 9 valores diferentes de *n* (*n*=2,3,...10), conforme será visto no Capítulo 4 – Testes e Resultados.

Ao final do processo de extração de características cada região de interesse é representada por um vetor de valores, cujo tamanho depende do tipo de característica usado, da abordagem de extração e do número de áreas escolhidas.

3.6 Seleção de Características e Classificação

O objetivo desta etapa da metodologia é utilizar os vetores de características extraídos das regiões de interesse na etapa anterior para treinar um classificador MVS e em seguida classificar essas regiões em massas ou não-massas. No entanto, conforme explicado na Seção 2.6, para que o classificador MVS atinja um bom poder de generalização e apresente resultados de classificação satisfatórios é necessário realizar o procedimento de seleção das características mais relevantes, ou seja, as que melhor discriminem as duas classes a serem diferenciadas.

Neste trabalho o procedimento de seleção de características e classificação das regiões de interesse é realizado em um processo híbrido AG-MVS tradicional (CHOW *et al.*, 2008), conforme esquema apresentado na Figura 21. Nesse processo, cada cromossomo é composto de uma seqüência binária de genes, indicando quais as características, cujos valores estão armazenados no vetor de características de cada região de interesse, que serão utilizadas durante o treinamento do classificador MVS. A busca genética é responsável por evoluir esses cromossomos para descobrir que subconjunto de características fornece ao classificador o maior poder de generalização durante o treinamento. Esse treinamento é realizado para cada cromossomo da população na forma de validação cruzada, ao final da qual são calculadas as medidas de desempenho obtidas pelo classificador MVS.

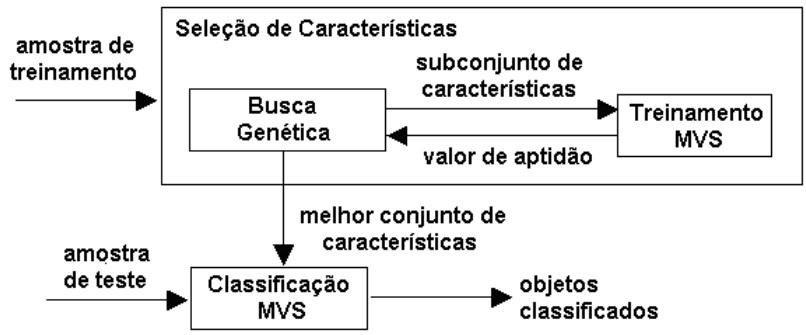

Figura 21 – Esquema do processo integrado de seleção de características e classificação das regiões de interesse.

Testes preliminares indicaram a medida *F-Measure*, como sendo a mais adequada para guiar o processo de evolução, em comparação com as demais medidas de desempenho apresentadas na Seção 2.7, o que levou a sua escolha como medida de aptidão dos cromossomos testados. Ou seja, quanto maior o valor de *F-Measure* obtido por um determinado cromossomo, maiores as chances das características representadas por ele se perpetuarem ao longo das gerações. A busca é interrompida quando o melhor cromossomo se perpetuar como mais apto por 100 gerações consecutivas.

Apenas uma parte das regiões de interesse, chamada de conjunto de treinamento, é utilizada no processo de seleção de características e treinamento do classificador. Os demais, chamados de conjunto de teste, são utilizados para verificar se o modelo gerado pelo MVS, durante o treinamento, de fato, possibilita uma boa classificação.

Neste trabalho, o classificador MVS foi utilizado com núcleo radial e parâmetros padrão (C=1 e y=0,5). O algoritmo genético foi implementado e utilizado com a seguinte configuração:

- População: 20 cromossomos inicializados aleatoriamente;
- Medida de aptidão: *F-Measure*;
- Seleção: roleta proporcional à medida de aptidão;
- Cruzamento: 1 ponto aleatório;
- Elitismo: 2 indivíduos mais aptos da geração são perpetuados;
- Porcentagem de mutação: 5% dos filhos gerados;

O próximo capítulo descreve os diversos experimentos realizados para testar a metodologia e discute os resultados obtidos.

4 TESTES E RESULTADOS

Para avaliar a metodologia de detecção de massas proposta neste trabalho uma série de testes foi realizada. Esta seção apresenta e discute os resultados obtidos nas diversas abordagens utilizadas.

A metodologia foi testada com uma amostra de 650 imagens extraídas do DDSM, cada uma contendo apenas uma massa. A partir dessas imagens a etapa de segmentação das regiões de interesse selecionou um total de 2679 regiões suspeitas de conterem anormalidades, sendo 603 massas de fato e de 2076 não massas. Em 47 imagens das 650 iniciais o procedimento de segmentação falhou em incluir as massas presentes no conjunto de regiões de interesse, o que representa 7,23% dos casos. As 603 massas segmentadas corretamente representam 92,77% dos casos. Esses resultados demonstram que a etapa de segmentação apresenta uma boa sensibilidade, mas produz muitos falsos positivos, mais precisamente 3,19 por imagem (2076 regiões que não contem massas, segmentadas como regiões de interesse nas 650 imagens), os quais se espera que sejam eliminados durante a etapa de classificação.

Durante a etapa de seleção de características e classificação as regiões de interesse foram divididas em grupos de treinamento e teste utilizando seis proporções diferentes: 30/70, 40/60, 50/50, 60/40, 70/30 e 80/20, o primeiro número indicando a porcentagem de regiões utilizada para treinamento e o segundo a porcentagem de regiões utilizada para teste. A seleção dos objetos em cada grupo foi feita aleatoriamente a partir do total de regiões segmentadas.

Diversas abordagens de testes foram utilizadas envolvendo as características de geometria e textura definidas nas Seções 2.3 e 2.4. Os resultados foram verificados de forma automática através da comparação da região detectada pela metodologia com a aquela indicada pelo especialista, de acordo com as informações do DDSM. As próximas seções descrevem cada uma dessas abordagens e discute os resultados obtidos.

4.1 Testes usando Geometria

Na primeira abordagem de testes, utilizaram-se apenas as características de geometria para descrever as regiões a serem classificadas. O vetor de características, portanto, foi formado pelos valores de excentricidade, circularidade, compacidade, desproporção circular e densidade circular de cada região, sendo o cromossomo codificado nessa mesma ordem, ou seja, uma seqüência de genes *01010* indica, por exemplo, que somente os valores de circularidade e desproporção circular foram utilizados.

A Tabela 1 mostra os indicadores de desempenho obtidos com essa abordagem. A primeira coluna indica a proporção dos grupos de treinamento (Tr) e testes (Te) utilizada. A segunda mostra o cromossomo com maior *fitness* ao final da busca genética. As demais colunas apresentam os indicadores de desempenho obtidos, apresentando em vermelho o pior resultado para cada indicador e em azul o melhor.

Tabela 1 – Resultados dos testes usando apenas características geométricas.

Tr/Te	Cromossomo (E,C,Co,D,Dc)	A (%)	S (%)	E(%)	FP/i	FN/i	Fm (%)
30/70	10010	78,64	60,52	83,91	0,55	0,39	72,22
40/60	11011	80,41	58,56	86,76	0,46	0,41	72,66
50/50	10011	79,10	62,25	84,01	0,55	0,38	73,13
60/40	10111	82,67	64,88	87,85	0,42	0,35	76,37
70/30	10011	80,72	58,56	87,16	0,44	0,41	72,86
80/20	10011	81,38	68,60	85,10	0,51	0,31	76,85
	Média	**80,49**	**62,23**	**85,80**	**0,49**	**0,38**	**74,01**

Conforme pode ser observado na Tabela 1, essa abordagem atingiu uma acurácia média razoável (80,49%) com uma boa especificidade (85%), deixando a desejar, no entanto no quesito sensibilidade (62,2%), o que a deixou com um valor relativamente baixo de *F-Measure.*

Pode-se verificar ainda que as características mais selecionadas na melhor classificação de cada proporção dessa abordagem foram a excentricidade e a desproporção circular, ambas presentes em todas as proporções, alem da densidade circular, ausente apenas uma vez.

4.2 Testes usando Textura

Esta seção descreve os resultados obtidos pela abordagem que utiliza apenas a informação de textura das regiões de interesse, descrita através do Índice de Diversidade de Simpson. Essa abordagem se divide em quatro outras de acordo com o tipo de extração realizada: extração global, extração em anéis, extração em círculos e mista. As próximas subseções descrevem cada uma dessas abordagens.

4.2.1 Abordagem Global

Esta abordagem de testes utiliza a estratégia de extração global do Índice de Simpson, definida na Seção 3.5.2.1, para descrever a textura das regiões de interesse. Nesse caso específico como há apenas um único valor para representar cada região, não há necessidade de realizar a seleção de características e os dados são passados diretamente para o classificador MVS. Os resultados obtidos por essa abordagem são apresentados na Tabela 2, seguindo a mesma estrutura da Tabela 1, com exceção da coluna Cromossomo, que não apresenta valores já que a busca genética não é realizada.

Tabela 2 – Resultado dos testes usando a abordagem de extração global do Índice de Diversidade de Simpson.

Tr/Te	Cromossomo	A(%)	S(%)	E(%)	FP/i	FN/i	Fm(%)
30/70	-	77,09	65,25	80,54	0,67	0,35	73,00
40/60	-	75,12	67,13	77,45	0,78	0,33	72,09
50/50	-	76,64	65,89	79,77	0,70	0,34	72,17
60/40	-	76,89	67,36	79,66	0,70	0,33	71,92
70/30	-	75,87	58,56	80,90	0,66	0,41	70,14
80/20	-	74,86	63,64	78,13	0,75	0,36	67,94
	Média	**76,08**	**64,64**	**79,41**	**0,71**	**0,35**	**71,21**

Os resultados apresentados na Tabela 2 demonstram que essa abordagem teve um desempenho pior que a abordagem baseada em

geometria, pois atingiu menores valores em todos os indicadores, exceto na sensibilidade, onde apresentou um pequeno ganho: 64,64% contra 62,23%.

4.2.2 Abordagem em Anéis

Esta abordagem utiliza a estratégia de extração por regiões em forma de anéis do Índice de Simpson. A fim de identificar o número ideal de anéis em que as regiões de interesse devem ser divididas, foram realizados testes com várias quantidades diferentes (*n*). Para cada valor de *n* uma bateria de testes utilizando as seis proporções de treinamento foi executada. Cada linha da Tabela 3 mostra a média dos resultados obtidos com as seis proporções para cada valor de *n.* Os cromossomos aqui foram codificados representando as áreas das regiões de interesse que serão selecionadas ou não pela busca genética, ordenadas do anel mais externo para o mais interno. Ou seja, para *n*=3, por exemplo, um cromossomo 001 indica que somente o anel mais interno da região de interesse foi selecionado.

Tabela 3 – Resultado dos testes usando a abordagem de extração em anéis do Índice de Diversidade de Simpson.

n	Cromossomo	A(%)	S(%)	E(%)	FP/i	FN/i	Fm
2	11	76,79	71,90	78,22	0,75	0,28	74,93
3	111	75,58	70,66	77,02	0,79	0,29	73,70
4	1011	75,96	70,25	77,62	0,77	0,30	73,75
5	11101	74,28	71,49	75,09	0,86	0,29	73,25
6	110011	74,75	69,50	76,27	0,82	0,30	72,73
7	1110011	75,07	68,32	77,03	0,79	0,32	72,41
8	01001110	75,40	70,25	76,90	0,79	0,30	73,42
9	110111011	74,37	70,66	75,45	0,84	0,29	72,98
10	1010000010	74,53	71,63	75,38	0,85	0,28	73,46
	Média	**75,19**	**70,52**	**76,55**	**0,81**	**0,29**	**73,40**

Comparando os resultados da Tabela 2 com os da Tabela 3, verifica-se que a abordagem de extração em anéis foi bem melhor que a abordagem de extração global no quesito sensibilidade, sendo apenas ligeiramente inferior nos quesitos acurácia e especificidade. Esses resultados demonstram que a informação de textura pode apresentar diferentes padrões em diferentes localizações das regiões de interesse. Comparando o caso específico da

abordagem por anéis utilizando dois anéis (*n*=2) com a abordagem global, observa-se que o classificador conseguiu discriminar melhor regiões de interesse que continham massas quando teve valores locais de textura, no caso, a textura do anel mais externo e a textura da área mais interna, fato evidenciado pelo aumento considerável no valor da sensibilidade.

Analisando os cromossomos apresentados na Tabela 3, observa-se que, à exceção do caso *n*=3, em geral, os anéis mais selecionados durante a busca genética são os mais externos e os mais internos, indicando possivelmente que as informações de textura das áreas intermediárias das regiões de interesse não são tão relevantes para a discriminação das massas.

4.2.3 Abordagem Circular

Esta abordagem utiliza a estratégia de extração por regiões em forma de círculos do Índice de Simpson e segue a mesma estrutura da abordagem anterior no que diz respeito às varias quantidades de regiões utilizadas à codificação dos cromossomos. A Tabela 4 mostra os indicadores de desempenho alcançados por essa abordagem.

Tabela 4 – Resultado dos testes usando a abordagem de extração circular do Índice de Diversidade de Simpson.

n	Cromossomo	A(%)	S(%)	E(%)	FP/i	FN/i	Fm(%)
2	11	78,84	74,79	80,02	0,69	0,25	77,32
3	111	79,12	76,03	80,02	0,69	0,24	77,97
4	1101	78,94	75,62	79,90	0,69	0,24	77,70
5	10101	78,66	76,86	79,18	0,71	0,23	78,00
6	101110	78,94	76,45	79,66	0,70	0,24	78,02
7	1010011	79,22	76,86	79,90	0,69	0,23	78,35
8	10010010	79,12	77,27	79,66	0,70	0,23	78,45
9	100110010	79,03	77,27	79,54	0,70	0,23	78,39
10	1001100010	79,12	77,27	79,66	0,70	0,23	78,45
	Média	**79,00**	**76,49**	**79,73**	**0,70**	**0,24**	**78,07**

Essa abordagem proporcionou melhores resultados que as abordagens de textura anteriores, sendo superior a elas em todos os indicadores de desempenho, obtendo uma média de 79% de acurácia, 76,49% de sensibilidade e 79,73% de especificidade. Diferentemente da abordagem por

anéis a observação dos cromossomos selecionados na abordagem circular indica que as regiões intermediarias também contribuem para a discriminação das massas.

4.2.4 Abordagem Mista

Essa abordagem combina as três abordagens de extração de textura anteriores, codificando o cromossomo com todas as medidas extraídas. É importante ressaltar, conforme exposto na Seção 3.5, que o círculo mais externo para um dado valor de *n* na extração circular corresponde à extração global e que a área mais interna das duas abordagens por região são coincidentes. Assim, essas informações não são repetidas na codificação do cromossomo, que é feita da seguinte maneira: os *n* primeiros genes representam as medidas extraídas através da abordagem circular para cada região, o que engloba, portanto, no gene zero, o valor global do Índice Simpson; os genes seguintes correspondem às medidas extraídas nos *n* anéis, com exceção do mais interno. Com isso obtém-se um cromossomo de tamanho (2*n)-1, representando quais as medidas que serão utilizadas durante o processo integrado de seleção de características e classificação dos candidatos. A Tabela 5 mostra os valores obtidos.

Tabela 5 – Resultado dos testes usando simultaneamente as três abordagens de extração de textura.

n	Cromossomo	A(%)	S(%)	E(%)	FP/i	FN/i	Fm(%)
2	110	78,84	74,79	80,02	0,69	0,25	77,32
3	11010	79,40	76,86	80,14	0,68	0,23	78,47
4	1001101	78,94	77,27	79,42	0,71	0,23	78,33
5	101001000	79,78	78,51	80,14	0,68	0,21	79,32
6	10100010000	79,14	79,34	79,09	0,72	0,21	79,21
7	1010000100000	79,68	79,75	79,66	0,70	0,20	79,70
8	100100001000000	79,87	79,34	80,02	0,69	0,21	79,68
9	11110000010000000	80,43	79,34	80,75	0,66	0,21	80,04
10	1001000000100000001	79,87	79,75	79,90	0,69	0,20	79,82
	Média	**79,55**	**78,33**	**79,90**	**0,69**	**0,22**	**79,10**

Os resultados obtidos pela abordagem conjunta foram superiores aos resultados obtidos por cada uma das outras abordagens de extração do Índice de Simpson individualmente, apresentando uma melhora geral dos indicadores, especialmente na sensibilidade.

O gráfico da Figura 22 permite observar com mais clareza a diferença de desempenho entre as abordagens que utilizam o Índice de Diversidade de Simpson. As abordagens por região, especialmente a abordagem circular, se mostraram mais eficientes em descrever a textura das massas, uma vez que a sensibilidade da classificação atinge um maior valor nessas abordagens. A abordagem conjunta propicia ainda um pequeno ganho em comparação com a abordagem circular, que foi a melhor nos testes individuais.

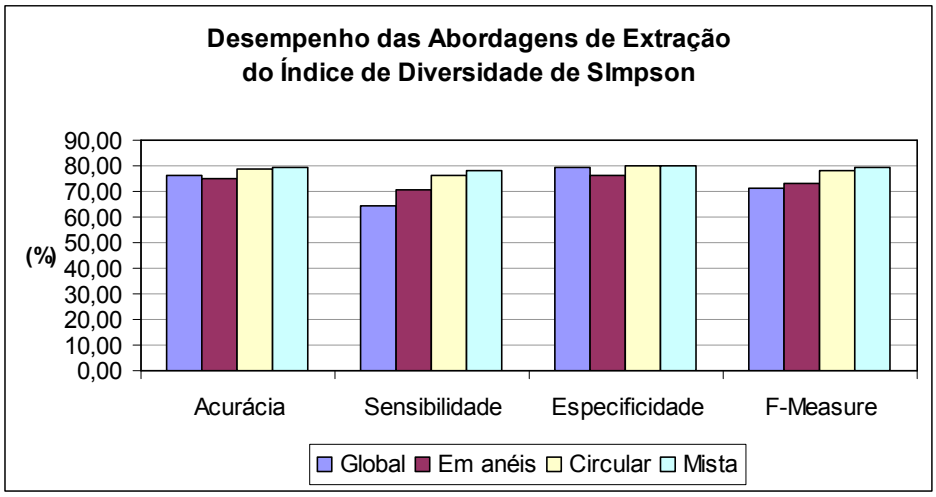

Figura 22 – Comparação entre as abordagens de extração de textura.

Analisando todos os testes apresentados até aqui, pode-se observar que o Índice de Diversidade de Simpson proporcionou resultados mais equilibrados que os obtidos com as medidas geométricas, no entanto em termos de especificidade a abordagem geométrica apresentou valores muito maiores.

4.3 Testes usando Geometria e Textura

Esta abordagem utiliza tanto as características de geometria quanto as informações de textura, extraídas de diversas formas. O objetivo é tentar fornecer ao classificador MVS os melhores aspectos dos dois tipos de medidas, qual sejam, a boa especificidade proporcionada pelas características de geometria e a boa sensibilidade fornecida pelas características de textura.

A codificação dos cromossomos se deu de forma semelhante à abordagem de textura mista, mas inserindo os cinco genes que representam as características geométricas antes dos genes que representam as características de textura, ou seja, o tamanho do cromossomo passa a ser 5+[(2*n)-1], sendo n a quantidade de anéis ou círculos em que as regiões de interesse são divididas. A Tabela 6 mostra os indicadores de desempenho obtidos.

Conforme esperado, as duas características proporcionaram juntas um maior poder de generalização ao classificador, atingindo em média 83,94% de acurácia, 83,24% de sensibilidade, e 84,14% de especificidade, com taxa média de falsos positivos por imagem e falsos negativos por imagem, 0,55 e 0,17, respectivamente. Esses resultados superam todos os outros resultados individuais em todos os indicadores de desempenho.

Tabela 6 – Resultado dos testes usando simultaneamente as características de geometria e de textura.

n	Cromossomo	A(%)	S(%)	E(%)	FP/i	FN/i	Fm(%)
2	11111110	83,60	82,64	83,87	0,55	0,17	83,25
3	1101111000	84,06	83,06	84,36	0,54	0,17	83,70
4	110111010001	83,97	82,64	84,36	0,54	0,17	83,49
5	10011100010010	83,60	83,88	83,51	0,57	0,16	83,69
6	1101110100000000	84,06	83,06	84,36	0,54	0,17	83,70
7	1101110100000000000	83,88	83,47	84,00	0,55	0,17	83,73
8	11011100000001000000	84,25	83,47	84,48	0,53	0,17	83,97
9	110111000100100000000	83,78	83,88	83,75	0,56	0,16	83,81
10	1101110010000000000000000	84,25	83,06	84,60	0,53	0,17	83,82
	Média	**83,94**	**83,24**	**84,14**	**0,55**	**0,17**	**83,69**

4.4 Resultado Final

Esta seção discute os principais resultados obtidos pela metodologia nas seis abordagens de testes utilizadas. A Tabela 7 mostra um resumo dos indicadores de desempenho médios alcançados por cada abordagem, durante a classificação MVS das regiões de interesse em massas e não massas.

Analisando os resultados observa-se que os testes utilizando apenas características geométricas apresentaram a melhor especificidade média (85,80%) e a melhor taxa de falsos positivo por imagem (0,49), deixando muito a desejar, entretanto, no que diz respeito à sensibilidade (62,23%) e à taxa de falsos negativos por imagem (0,38), onde obteve os piores valores entre todas as abordagens testadas. Esses resultados indicam que as características geométricas possuem um bom poder discriminatório dos tecidos normais da mama sem, contudo, apresentar índices satisfatórios para a discriminação das anormalidades.

Tabela 7 – Resumo dos resultados obtidos pela metodologia.

Abordagem	A(%)	S(%)	E(%)	FP/i	FN/i	Fm(%)
Geometria	80,49	62,23	85,80	0,49	0,38	74,01
Simpson – Extração Global	76,08	64,64	79,41	0,71	0,35	71,21
Simpson – Extração em Anéis	75,19	70,52	76,55	0,81	0,29	73,40
Simpson – Extração Circular	79,00	76,49	79,73	0,70	0,24	78,07
Simpson – Abordagem Mista	79,55	78,33	79,90	0,69	0,22	79,10
Geometria e Simpson	83,94	83,24	84,14	0,55	0,17	83,69

As abordagens que utilizam o Índice de Diversidade de Simpson, especialmente a abordagem circular e a abordagem mista, apresentaram um ganho significativo na sensibilidade da classificação em relação à abordagem geométrica. No entanto em relação à especificidade, houve perdas de mais de 5 pontos percentuais, indicando que a abordagem de textura utilizada é melhor para a discriminação das massas.

A utilização conjunta das abordagens de geometria e textura apresentou o resultado mais robusto quanto à capacidade de discriminação de massas e não massas, pois foi a melhor em todos os indicadores de desempenho, exceto na média de falsos positivos por imagem, onde ainda foi pior que a abordagem

usando apenas geometria. A utilização dos dois tipos de características juntas proporcionou ao classificador os melhores aspectos de cada uma delas: o poder de discriminação de massas proporcionado pelo Índice de Diversidade de Simpson e o poder de discriminação de não massas fornecido pelas características geométricas.

A comparação entre as três abordagens básicas de extração do Índice de Simpson – global, em anéis e circular – demonstrou um aspecto interessante da textura das massas. Ao que tudo indica a distribuição dos *pixels* nas regiões próximas às bordas das massas apresentam um padrão de diversidade característico, distinto do padrão apresentado pelas regiões mais internas e pela massa como um todo. A conclusão é baseada no fato de que o classificador atingiu uma sensibilidade muito melhor utilizando as abordagens de extração por regiões em comparação com a abordagem global. Essa informação local de textura, portanto, é uma característica importante e merece estudos mais aprofundados, para que se possa melhorar o desempenho desta metodologia.

4.5 Estudos de Casos

Esta seção examina as etapas mais importantes da metodologia proposta a partir de alguns casos de testes reais. O objetivo é facilitar a compreensão das técnicas utilizadas, e do fluxo de processamento como um todo, através das imagens geradas por cada etapa. Para isso serão examinados três casos. O primeiro caso é um exemplo em que a metodologia obteve êxito total na detecção da massa, ou seja, conseguiu uma boa segmentação das regiões de interesse e uma classificação correta dessas regiões. O segundo caso examinado mostra um exemplo em que a metodologia também realizou uma boa segmentação, mas falhou em classificar corretamente as regiões segmentadas. O terceiro caso apresenta uma situação em que a metodologia não obteve êxito em segmentar adequadamente as regiões de interesse, comprometendo o resultado final apesar de a classificação subseqüente ter sido realizada a contento.

4.5.1 Detecção Correta

O primeiro caso, apresentado na Figura 23, mostra a seqüência de passos realizados para a detecção correta de uma massa em uma imagem de mamografia. A Figura 23-a mostra a imagem original da mamografia em questão, tal como se apresenta na base do DDSM.

O primeiro passo da metodologia, o pré-processamento, produz como resultado a imagem da Figura 23-b, na qual se pode observar que os objetos indesejáveis, como o fundo e as marcas externas a mama, foram removidos. Também é possível observar os efeitos do realce linear de contraste realizado na discriminação visual das estruturas mais densas da mama em relação às menos densas.

A próxima etapa realiza a segmentação das regiões de interesse e divide-se em dois passos subseqüentes. Primeiro a imagem pré-processada da Figura 2-b é submetida ao algoritmo de agrupamento *K-Means*, em seguida os grupos gerados são percorridos pela técnica de *Template Matching.* A Figura 23-c mostra os agrupamentos produzidos para cada quantidade de grupos *k* utilizada. Vários valores de *k* são utilizados porque não se sabe *a priori* qual a quantidade de grupos mais adequada para segmentar as massas em cada imagem testada. A Figura 23-d mostra a coleção de regiões de interesse resultante da varredura dos três *templates* utilizados sobre as estruturas isoladas em cada agrupamento da Figura 23-c.

A seguir tem-se a etapa de extração de características das regiões de interesse segmentadas, a qual não produz resultados visuais, e sim um vetor de valores representando as características extraídas. A etapa seguinte utiliza esse vetor de características para classificar as regiões de interesse em massa e não massa através de uma MVS previamente treinada. A Figura 23-e apresenta, em azul, as regiões classificadas como massas – nesse caso especifico, apenas uma. Para verificar o êxito da detecção, a marcação da localização correta da massa, obtida a partir das informações contidas no DDSM, é impressa sobre a imagem resultante. A Figura 23-e permite observar que no caso examinado a detecção realizada pela metodologia foi extremamente precisa.

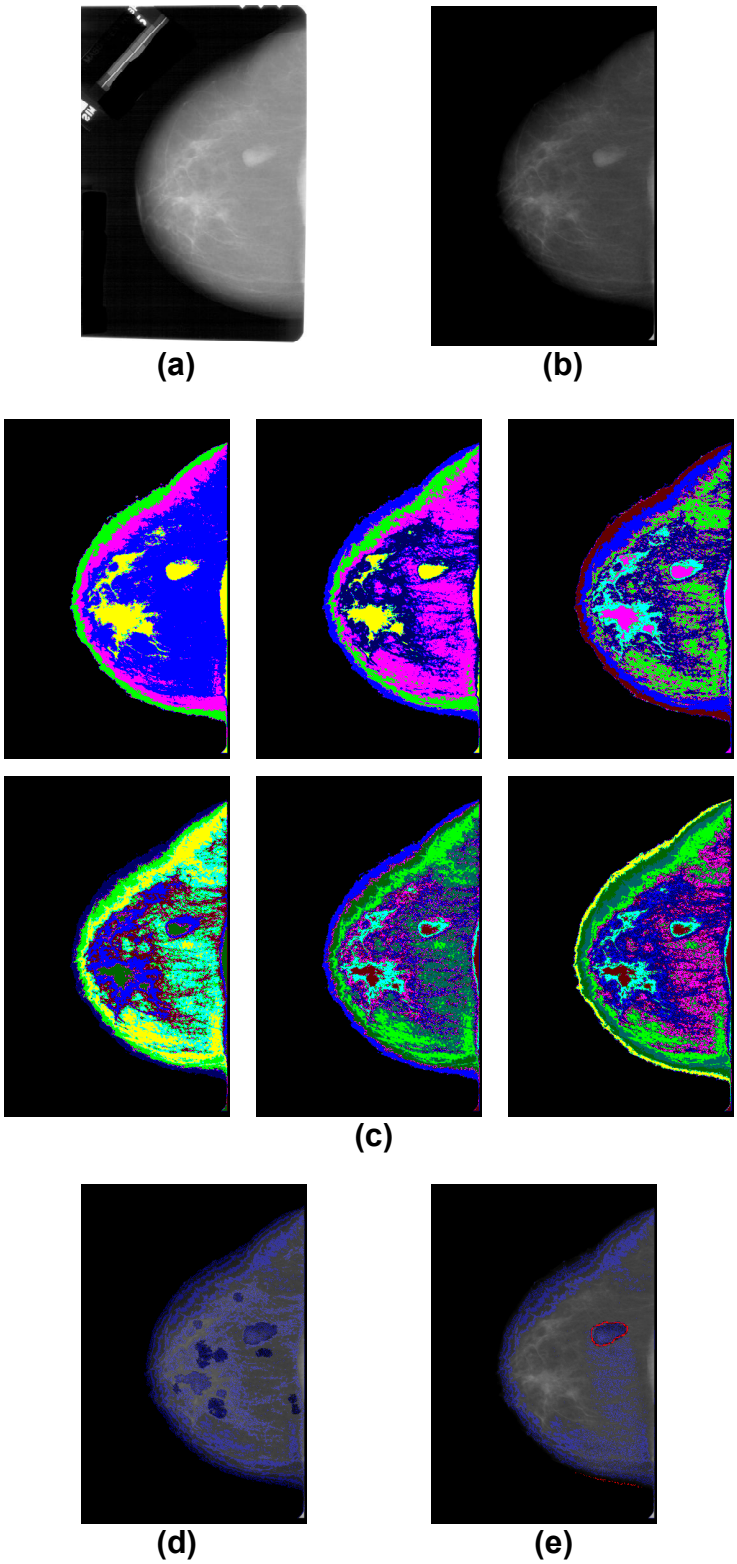

Figura 23 – Imagens do Estudo de Caso 1: (a) Original; (b) Pré-processada; (c) Agrupamentos gerados pelo *K-Means* (*k*=5,6,7,8,9,10); (d) Regiões de Interesse selecionadas pelo *Template Matching*; (e) Região classificada como massa pela MVS. Nesse caso a metodologia foi bem sucedida em detectar a massa, segundo informações do DDSM (marcação em vermelho).

4.5.2 Falha da Classificação

O segundo caso, apresentado na Figura 24, mostra a mesma seqüência de passos descritos na seção anterior: imagem original (Figura 24-a); resultado do pré-processamento (Figura 24-b); agrupamentos gerados (Figura 24-c); regiões de interesse segmentadas (Figura 24-d) e o resultado da classificação, mostrando em azul a região classificada pela MVS como massa (Figura 24-e). Como a marcação em vermelho indica a localização correta da massa, obtida a partir das informações disponíveis no DDSM, pode-se observar na Figura 24-e que, apesar da massa ter sido segmentada entre as regiões de interesse, o classificador não obteve êxito em classificá-la adequadamente, descartando-a e apresentando como massa uma região que, na verdade, corresponde a um tecido normal. Esse caso ilustra a ocorrência de um falso negativo (tecido anormal classificado como normal) e um falso positivo (tecido normal classificado como anormal). Entre os possíveis fatores relacionados a essa falha está o fato de que a mama deste exemplo apresenta uma densidade muito alta em quase toda a sua extensão, o que faz com que as bordas das massas sejam pouco distinguíveis dos tecidos normais adjacentes, dificultando a descriminação de textura e induzindo o classificador a erros.

4.5.3 Falha da Segmentação

O terceiro caso apresenta um exemplo de falha da metodologia em realizar uma segmentação adequada. A Figura 25 exibe as imagens obtidas durante o processamento, seguindo a mesma ordem dos casos anteriores. Nesse caso, a etapa de segmentação falhou em incluir a massa entre as regiões de interesse (Figura 25-d), conforme se pode observar pela marcação da localização correta da massa, em vermelho, na Figura 25-e. Assim, mesmo que a etapa de classificação tenha sido eficiente em classificar todas as regiões segmentadas como tecidos normais, o resultado final foi comprometido, ocasionando um falso negativo. A possível causa da falha de segmentação, especula-se, também está relacionada à alta densidade dos tecidos da mama em questão, impedindo que o *K-Means* conseguisse diferenciar a intensidade dos *pixels* da massa em um agrupamento isolado.

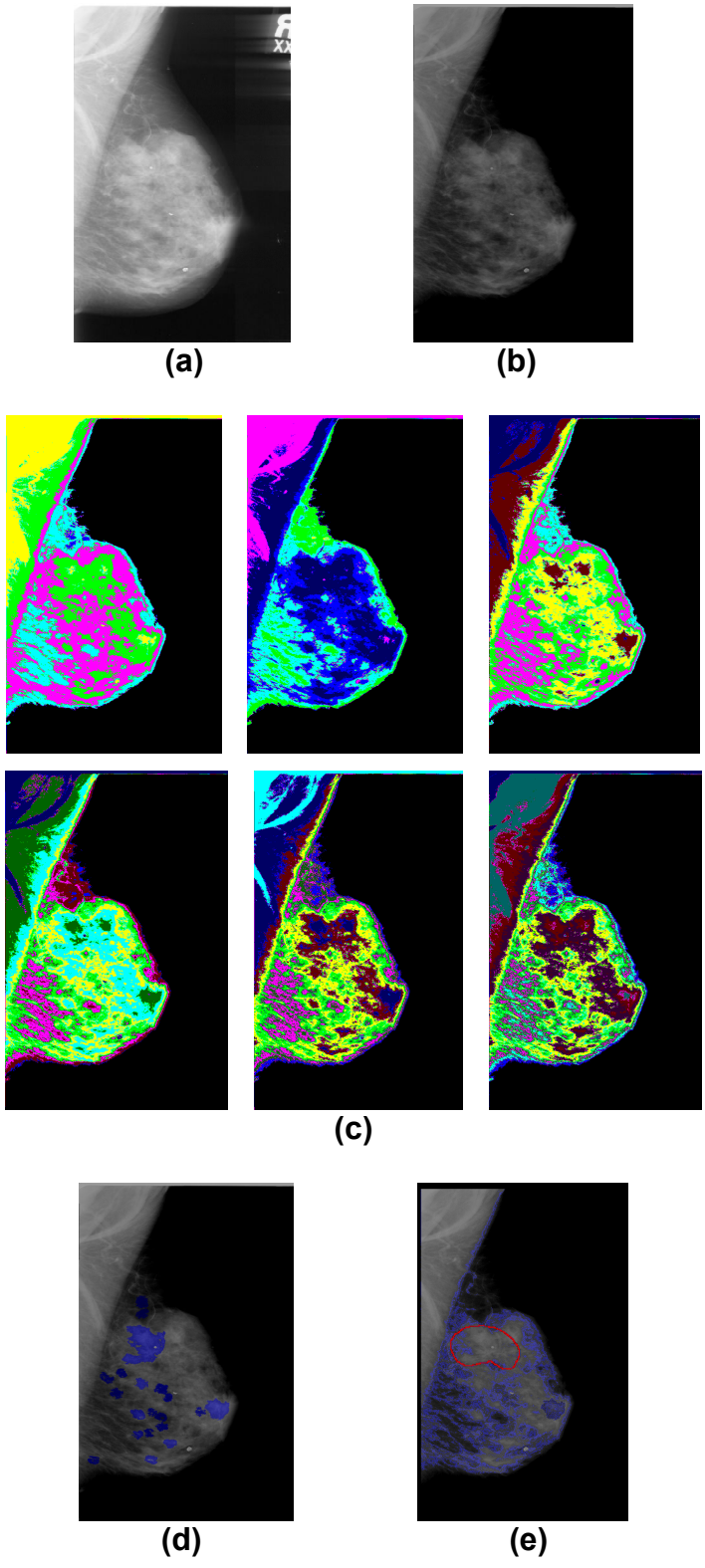

Figura 24 – Imagens do Estudo de Caso 2: (a) Original; (b) Pré-processada; (c) Agrupamentos gerados pelo *K-Means* (*k*=5,6,7,8,9,10); (d) Regiões de Interesse selecionadas pelo *Template Matching*; (e) Região classificada como massa pela MVS. Nesse caso a metodologia falhou em detectar a massa, segundo informações do DDSM (marcação em vermelho).

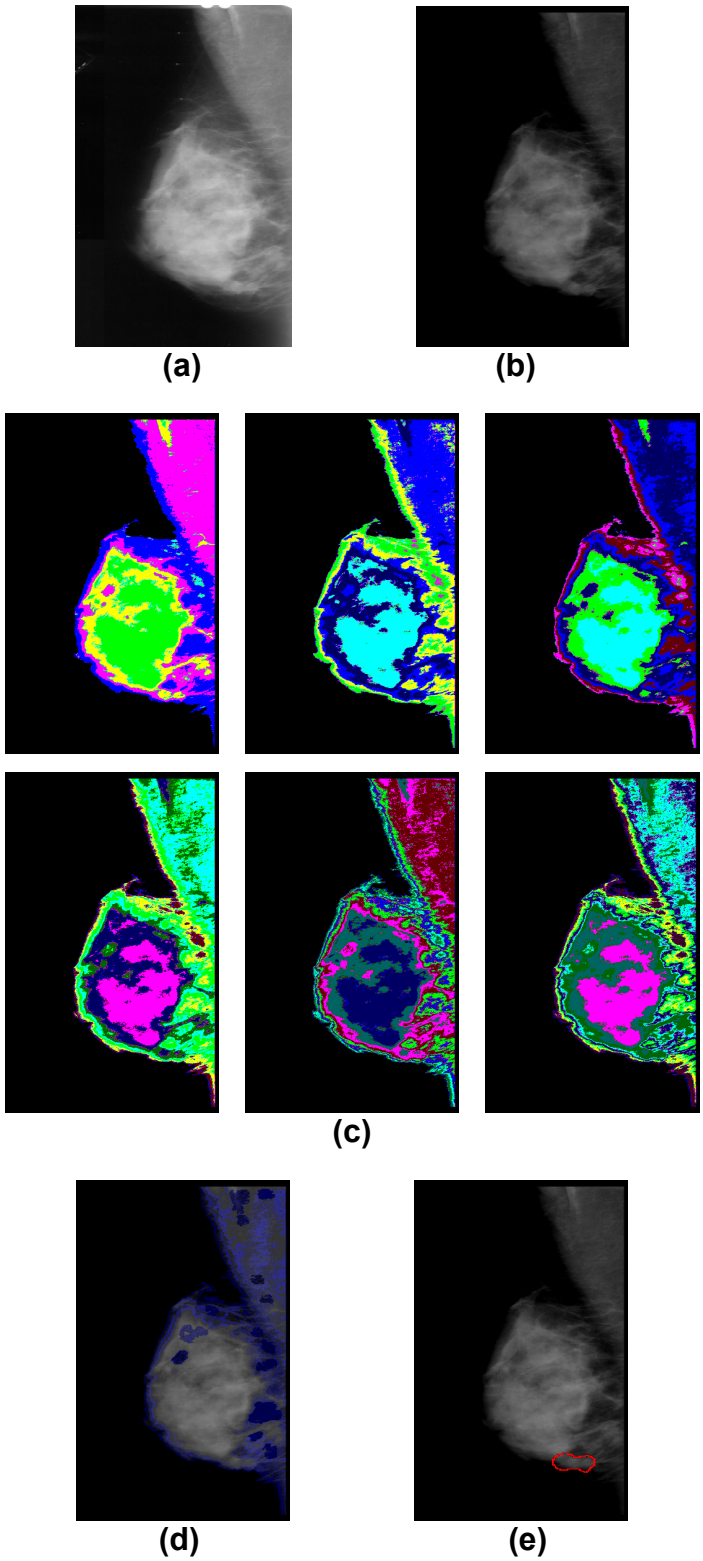

Figura 25 – Imagens do Estudo de Caso 3: (a) Original; (b) Pré-processada; (c) Agrupamentos gerados pelo *K-Means* (*k*=5,6,7,8,9,10); (d) Regiões de Interesse selecionadas pelo *Template Matching*; (e) Resultado da classificação MVS. Nesse caso a metodologia falhou em segmentar a região contendo a massa entre as regiões de interesse.

5 CONCLUSÃO

A elevada taxa de incidência e mortes causadas pelo câncer de mama, atualmente, no Brasil e no mundo, justifica o desenvolvimento de pesquisas científicas voltadas para estratégias de auxílio na detecção precoce da doença, fator determinante para o sucesso do tratamento.

O uso de ferramentas computacionais para auxílio à detecção e diagnóstico tem crescido em aceitação nos últimos anos, fornecendo uma segunda opinião para os especialistas em análise de imagens médicas, que cada vez mais as utilizam no seu dia a dia.

Este trabalho apresentou uma metodologia CAD para a detecção de massas em imagens digitais de mamografia, utilizando o algoritmo de agrupamento *K-Means* e a técnica de *Template Matching* para segmentação das regiões de interesse, as quais são posteriormente classificadas em massas ou não massas por um classificador MVS, previamente treinado para reconhecer os padrões de geometria e textura das duas classes.

Os resultados apresentados no Capítulo 4 evidenciaram o desempenho promissor da metodologia desenvolvida e a eficiência das técnicas utilizadas nas diversas etapas do processamento. A etapa de segmentação das regiões de interesse conseguiu segmentar 603 das 650 massas da amostra, o que equivale a 92,77% dos casos. A etapa de classificação das regiões segmentadas também obteve um desempenho aceitável, atingindo, em média, 83,94% de acurácia, 83,24% de sensibilidade, e 84,14% de especificidade, com taxa média de falsos positivos por imagem e falsos negativos por imagem de 0,55 e 0,17, respectivamente. Tais resultados indicam que o Índice de Diversidade de Simpson é uma medida promissora para caracterização da textura em imagens radiológicas, encorajando estudos mais aprofundados sobre a utilização desse tipo de medida em problemas de classificação de massas e não-massas através de MVS.

Entretanto, apesar dos bons resultados obtidos, diversos aspectos da metodologia podem ser melhorados possibilitando resultados ainda melhores. Um desses aspectos, por exemplo, é a forma como a etapa de segmentação

das regiões de interesse foi estruturada. Os diversos agrupamentos realizados sobre as imagens pré-processadas e os três tamanhos de *templates* utilizados, mesmo após a eliminação das regiões sobrepostas, resultam na segmentação de um número muito grande de regiões que não contém massas em comparação com a quantidade de regiões que contém massas. Essa desproporção numérica entre as duas classes acaba por influenciar o classificador, já que dispõe de muito mais informações sobre uma classe que de outra, fato evidenciado pela superioridade dos valores de especificidade em comparação com os valores de sensibilidade alcançados. A extensão dessa influência precisa ser avaliada para que se verifique a necessidade de se aperfeiçoar a etapa de segmentação para que produza uma quantidade mais balanceada de regiões suspeitas.

Outro aspecto que pode ser melhor explorado é o esquema de seleção de características baseado em busca genética. A realização de testes mais abrangentes no que diz respeito à escolha do indicador de desempenho mais adequado para guiar o processo de evolução tornaria o método mais robusto e consistente. Além disso, análises estatísticas sobre as características individuais mais selecionadas no processo nos dariam a noção exata de onde trabalhar para melhorar a etapa de seleção de características, proporcionando ao classificador MVS um maior poder de discriminação.

Além desses aspectos, diversas outras idéias surgiram ao longo do desenvolvimento deste trabalho, mas não puderam ser concluídas e inclusas, deixando algumas possibilidades em aberto a para trabalhos futuros. Entre elas estão: a pesquisa de outras medidas geométricas para caracterização das massas; a utilização do Índice de Diversidade de Simpson para classificar as massas detectadas de acordo com suas naturezas malignas ou benignas; a utilização de um outro classificador para comparação com os resultados obtidos pela MVS; a utilização de outros índices de diversidade para comparação com os resultados alcançados pelo Índice de Diversidade de Simpson; e a realização de estudos mais aprofundados sobre os padrões de diversidade locais apresentados pelas diferentes regiões das massas.

Por fim, o presente trabalho abre a possibilidade para utilização do Índice de Diversidade de Simpson para a descrição da textura de outros tipos de lesões como calcificações da mama, nódulos pulmonares, etc.

REFERÊNCIAS

BRAZ JÚNIOR, G., SILVA, E. C., PAIVA, A. C., SILVA, A. C., GATTASS, M., 2007. *Breast Tissues Mammograms Images Classification using Moran s Index, Geary s Coefficient and SVM*. In: International Conference on Neural Information Processing, 2007, Kitakyushu. Lecture Notes in Computer Science, LNCS, 2007.

BURHENNE, L. J. W., WOOD, S. A., D'ORSI, C. J., *et al*. The potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000; 215:554–562.

BUSHBERG, J. T., SEIBERT, J., A., LEIDHOLDT JR, E., M., BOONE, J. M., 2002. *The Essential Physics Of Medical Imaging*. Lippincott Williams & Wilkins, 2nd Edition. Philadelphia.

CAMARA, G., SOUZA, R. C. M., FREITAS, U. M., GARRIDO, J. C. P., 1996. *Integrating remote sensing and GIS by object-oriented data modeling.* Comput. Graph, v.20, n.3, 1996.

CAMPOS, L. F. A., SILVA, A. C., BARROS, A. K., 2007. *Independent Component Analysis and Neural Networks Applied for Classification of Malignant, Benign and Normal Tissue in Digital Mammography*. Methods of Information in Medicine, v. 46, p. 212-215, 2007

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

CHAVES, A. C. F., 2006. *Extração de Regras Fuzzy para Máquinas de Vetor de Suporte (SVM) para Classificação em Múltiplas Classes*. PhD Thesis. Pontifícia Universidade Católica do Rio de Janeiro.

CHOW, R., ZHONG, W., BLACKMON, M., STOLZ, R., DOWELL, M., 2008. *An efficient SVM-GA feature selection model for large healthcare databases.* Proceedings of the 10th annual conference on Genetic and evolutionary computation, July 12-16, 2008, Atlanta, GA, USA.

CRISTIANINI, N. e SHAWE-TAYLOR, J., 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

COSTA, D. D., BARROS, A. K., e SILVA, A. C., 2007. *Independent Component Analysis in Breast Tissues Mammograms Images Classification using LDA and SVM*. Information Technology Applications in Biomedicine - ITAB2007 - Tokyo. Conference on 6th International Special Topic pp. 231–234.

DDSM, 2001. *The Digital Database for Screening Mammography*, Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore and W. Philip Kegelmeyer, in Proceedings of the Fifth International Workshop on Digital Mammography, M.J. Yaffe, ed., 212-218, Medical Physics Publishing.

FENTON, J. J., TAPLIN, S. H., CARNEY, P. A., ABRAHAM, L., SICKLES, E. A., D'ORSI, C., BERNS, E. A., CUTTER, G., HENDRICK, R. E., BARLOW, W. E., ELMORE, J. G., 2007. *Influence of Computer-Aided Detection on Performance of Screening Mammography*. Breast Diseases: A Year Book Quarterly 18(3), 248-248.

FREER, T. W. e ULISSEY, M. J., 2001. *Screening Mammography with Computer-Aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center*. Radiology 220(3), 781–786.

GIGER, M. L, 2000. *Computer-aided diagnosis of breast lesions in medical images.* Computing in Science & Engineering, v. 2, n. 5, p. 39-45.

GOLDBERD, D., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. EUA: Addison-Wesley.

GONZALEZ, R. C. e WOODS, R. E., 2007. *Digital Image Processing*. 3rd Edition. Prentice Hall.

HARALICK, R. M., SHANMUGAN, K., DINSTEIN, I., 1973. *Texture Features for Image Classification*. IEEE Transactions on Systems, man and Cybernetics, v. 3, n. 6, p. 610-621, 1973.

HARTIGAN, J. e WANG, M., 1979. *A K-means clustering algorithm*. Applied Statistics, 28, 100–108.

HAYKIN, S. e ENGEL, P. M., 2001. *Redes Neurais: Principios e Pratica*. Bookman.

HEATH, M., BOWYER, K., KOPANS, D., MOORE, R., KEGELMEYER, W.P., 1998. *Current Status of the Digital Database for Screening Mammography.* Digital Mammography pp. 457–460.

iCad (2008). iCad Solutions. Disponível em: http://www.icadmed.com/.

INCA, 2008. Instituto Nacional do Câncer. *Estimativas 2008: Incidência de Câncer no Brasil*. Available at http://www.inca.gov.br.

INTEL, 2008. *Opencv, Open Computer Vision Library*. Intel Technology and Research. Disponível em http://sourceforge.net/projects/opencvlibrary/.

JAIN, A.K., MURTY, M.N., FLYNN, P.J., 1999. *Data Clustering: A Review*, ACM Comp. Surv.

JAIN, A. K., DUIN, R. P. W., MAO, J., 2000. *Statistical pattern recognition: A review*. IEEE Transactions on Pattern Analysis e Machine Intelligence, 22(1):4–37.

JENSEN, J. R., 1986. *Introductory digital image processing: a remote sensing perspective*. New Jersey: Prentice-Hall. 1986. p 379.

KOLLER, D., e SAHAMI, M., 1996. *Toward optimal feature selection*. In International Conference on Machine Learning, pages 284–292.

KOPANS, D. B., 2000. *Imagem da Mama*. Ed. MEDSI. Porto Alegre.

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

MAMOWEB, 2007. Projeto do Laboratório de Análise e Processamento de Imagens Médicas e Odontológicas. Escola de Engenharia de São Carlos. Disponível em: http://lapimo.sel.eesc.usp.br/lapimo/lapimo.htm

MARTINS, L. O., 2007. *Detecção de Massas em Imagens Mamográficas Através do Algoritmo Growing Neural Gas e da Função K De Ripley.* Dissertação de mestrado. Universidade Federal do Maranhão, Departamento de Engenharia de Eletricidade, Programa de Pós-Graduação em Engenharia de Eletricidade. São Luís, 2007.

MINISTÉRIO DA SAUDE, 2002. *Falando Sobre Câncer de Mama.* Instituto Nacional de Câncer, Coordenação de Prevenção e Vigilância. Rio de Janeiro.

MOAYEDI, F., BOOSTANI, R. AZIMIFAR, Z. KATEBI, S., 2007. *A Support Vector Based Fuzzy Neural Network Approach for Mass Classification in*

Mammography. Digital Signal Processing, 2007. 15th International Conference on pp. 240–243.

OSTA, H., QAHWAJI, R. e IPSON, S., 2008. *Wavelet-based Feature Extraction and Classification for Mammogram Images using RBF and SVM.* In Proceedings of International Conference on Visualization, Imaging, and Image Processing (VIIP). September 1 – 3, 2008. Palma de Mallorca, Spain.

PADWAL, M., 2007. *Elements of breasting imaging basics*. Disponivel em: http://www.gehealthcare.com/usen/ultrasound/education/products/cme_breast.html.

PAL, N. R. e PAL, S. K., 1993. *A Review on Image Segmentation Techniques.* Pattern Recognition, 26, 9, 1993, 1277–1294.

RANGAYYAN, R. M., EL-FARAMAWY, N. M., DESAUTELS, J. E. L., ALIM, O. A., 1997. *Measures of Acutance and Shape for Classification of Breast Tumors*. IEEE Transactions on Medical Imaging, V. 16, N. 6, 1997 p.799.

ROEHRIG, J., DOI, T., HASEGAWA, A., HUNT, B., MARSHALL, J., ROMSDAHL, H., SCHNEIDER, A., SHARBAUGH, R., ZANG, W., 1998. *Clinical Results with R2 ImageChecker in Support of FDA PMA Application*. *Fortschr Röntgenstr* 168, 175.

SAMPAT, M. P., MARKEY, M. K., BOVIK, A. C., 2005. *Computer-Aided Detection and Diagnosis in Mammography*. Handbook of Image and Vídeo Processing pp. 1195-1217.

SCHOUTEN, T., 2003. *Image Processing*. Radboud University, Department of Computer Science. 2003.

SIMPSON, E., 1949. Measurement of diversity. Nature, 163:688.

SOUSA, J. R., SILVA, A. C., PAIVA, A. C., 2007. *Lung Structures Classification Using 3D Geometric Measurements and SVM*. In: 12th Iberoamerican Congress on Pattern Recognition - CIARP 2007, Valparaiso. Lecture Notes Computer Science - LNCS. Berlin: Springer-Verlag, 2007. v. 4756. p. 783-792.

CAMARA G., SOUZA, R. C. M., FREITAS, U. M., GARRIDO, J., 1996: *Integrating remote sensing and GIS by object-oriented data modelling*. Computers & Graphics, 20: (3) 395-403, May-Jun 1996.

TIMP, S., VARELA, C., KARSSEMEIJER, N., 2007. *Temporal Change Analysis for Characterization of Mass Lesions in Mammography*. Medical Imaging, IEEE Transactions on 26(7), 945-953.

TUCERYAN, M. e JAIN, A., 1998. *Texture Analysis In The Handbook of Pattern Recognition and Computer Vision*, 207-248, World Scientific Publishing.

VAPNIK, V., 1998. Statistical Learning Theory. Wiley New York.

BURHENNE, W. L. J., WOOD, S. A., D'ORSI, C. J., *et al*., 2000. *Potential contribution of computer-aided detection to the sensitivity of screening mammography*. Radiology 2000; 215:554-562.

XING, E. P., 2003. *Feature selection in microarray analysis*. In Berrar, D., Dubitzky, W., e Granzow, M., editors, Understanding e Using Microarray Analysis Techniques: A Practical Guide, pages 110–131, Boston/Dordrecht/London. Kluwer Academic Publishers.

ZHANG, P. e KUMAR, K., 2006. *Analyzing Feature Significance from Various Systems for Mass Diagnosis*. Proceedings of the International Conference

on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce pp. 141-141.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo