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

WENER BORGES DE SAMPAIO

DETECÇÃO DE MASSAS EM IMAGENS MAMOGRÁFICAS USANDO REDES NEURAIS CELULARES, FUNÇÕES GEOESTATÍSTICAS E MÁQUINAS DE VETORES DE SUPORTE

São Luís
2009

Wener Borges de Sampaio

DETECÇÃO DE MASSAS EM IMAGENS MAMOGRÁFICAS USANDO REDES NEURAIS CELULARES, FUNÇÕES GEOESTATÍSTICAS E MÁQUINAS DE VETORES DE SUPORTE

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

Orientador: Prof. Dr. Aristófanes Corrêa Silva

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

São Luís
2009

Sampaio, Wener Borges de
Detecção de massas em imagens mamográficas usando redes neurais celulares, funções geoestatísticas e máquinas de vetores de suporte/ Wener Borges de Sampaio. – São Luís, 2009.
120 f.

Orientador: Prof. Dr. Aristófanes Corrêa Silva
Dissertação (Mestrado) – Curso de Pós Graduação em Engenharia de Eletricidade, Universidade Federal do Maranhão, 2009.
1. Mamografia. 2. Máquinas de Vetores de Suporte. 3. Redes Neurais Celulares. 4. Imagens mamográficas. I. Título.
CDU 618.19-073

DETECÇÃO DE MASSAS EM IMAGENS MAMOGRÁFICAS USANDO REDES NEURAIS CELULARES, FUNÇÕES GEOESTATÍSTICAS E MÁQUINAS DE VETORES DE SUPORTE

Wener Borges de Sampaio

Dissertação aprovada em 31 de agosto de 2009.

Prof. Aristófanes Corrêa Silva, Dr.
(Orientador)

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

Prof. Herman Martins Gomes, Dr.
(Membro da Banca Examinadora)

Profa. Alcione Miranda dos Santos, Dra.
(Membro da Banca Examinadora)

"Se enxerguei mais longe foi por estar sentado aos ombros de gigantes."
Isaac Newton

AGRADECIMENTOS

Agradeço a Deus por estar ao meu lado muito antes de minha existência e por ter colocado tantas pessoas importantes me acompanhando durante a longa caminhada de minha vida.

Aos meus pais, por todo sacrifício, carinho, educação, atenção, esperança e amor em mim depositados. Essa vitória também é de vocês.

Aos meus orientadores Aristófanes Silva e Anselmo Paiva por todo conhecimento a mim confiado, pela atenção e paciência que tanto exigi.

Aos professores Alexandre César, Allan Kardec, Denivaldo Lopes e Zair Abdelouahab que também tiveram parte importante na minha formação.

À minha noiva, Karla Teixeira, pelas lágrimas derramadas, pela força e incentivo, sem os quais não teria chegado aqui.

Aos amigos Alcides, André, Clériston, Edgar, Geraldo, Jocielma, Stelmo, Stefano pelo companheirismo, amizade e ajuda nas horas necessárias.

A todos que compõem a família UFMA-PPGEE, professores, alunos e funcionários, por manterem tão grandiosa estrutura de forma competente.

À Capes por financiar meus estudos nesses dois anos e permitir a viabilidade deste projeto.

A todos que oraram, torceram e incentivaram. Meu muito obrigado.

E em especial a Walleson por ser um amigo e irmão; Beatriz e família por terem me adotado nesses dois últimos anos.

RESUMO

Câncer de mama apresenta alta freqüência de ocorrência entre a população mundial e seus efeitos psicológicos alteram a percepção da sexualidade do paciente e a própria imagem pessoal. A mamografia é uma radiografia da mama que permite a descoberta precoce de câncer, sendo capaz a mostrar lesões nas fases iniciais, tipicamente lesões muito pequenas na ordem de milímetros. O processamento de imagens mamográficas tem contribuído para a descoberta e o diagnóstico de nódulos mamários, sendo uma importante ferramenta, pois reduz o grau de incerteza do diagnóstico, provendo uma fonte adicional de informação ao especialista. Este trabalho apresenta uma metodologia computacional que ajuda o especialista na descoberta de massas mamárias. O primeiro passo da metodologia visa à melhoria da imagem da mamografia que consiste em remoção de objetos externos à mama, redução de ruídos e realce das estruturas internas da mama. Então, Redes Neurais Celulares são usadas para segmentar áreas suspeitadas de conter massas. Estas regiões têm as suas formas analisadas por descritores de geometria (excentricidade, circularidade, densidade, desproporção circular e densidade circular) e as suas texturas analisadas por funções geoestatísticas (função de K de Ripley, e os índices de Moran e Geary). Máquinas de Vetores de Suporte são treinadas para classificar as regiões candidatas em um das classes, massas ou não-massa, com sensibilidade de 80,00%, especificidade de 85,68%, acurácia de 84,62%, uma taxa de 0,84 falsos positivos por imagem e 0,20 falsos negativos por imagem e uma área sob da curva ROC de 0,870.

Palavras-chave: Mamografia, Detecção Auxiliada por Computador, Redes Celulares Neurais, Índice de Moran, Coeficiente de Geary, Função K de Ripley, Máquina de Vetores de Suporte.

ABSTRACT

Breast cancer presents high occurrence frequency among the world population and its psychological effects alter the perception of the patient's sexuality and the own personal image. Mammography is an x-ray of the mamma that allows the precocious detection of cancer, since it is capable to showing lesions in their initial stages, typically very small lesions in the order of millimeters. The processing of mammographic images has been contributing to the detection and the diagnosis of mammary nodules, being an important tool, because it reduces the degree of uncertainty of the diagnosis, providing a supplementary source of information to the specialist. This work presents a computational methodology that aids the specialist in the detection of breast masses. The first step of the methodology aims at improvement the mammographic image, which consists of removal of unwanted objects, reduction of noise and enhancement of the breast internal structures. Then, Cellular Neural Networks are used to segment areas suspected of containing masses. These regions have their shapes analyzed by geometry descriptors (eccentricity, circularity, compactness, circular disproportion and circular density) and their textures are analyzed using geostatistical functions (Ripley's K function, Moran's and Geary's indices). Support Vector Machine were trained and used to classify the candidate regions in one of the classes, masses or no-mass, with sensibility of 80.00%, specificity of 85.68%, acuracy of 84.62%, a rate of 0.84 false positive for image and 0.20 false negative for image and an area under the curve ROC of 0.827.

Keywords: Mammography, Computer-Aided Detection, Cellular Neural Networks, Moran's Index, Geary's Index, Ripley's K Function, Support Vector Machine.

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE SIGLAS E ABREVIATURAS

ADL	Análise Discriminante Linear
ANCE	*Adaptive Neighborhood Contrast Enhancement* (Aumento de contraste adaptado à vizinhança)
AUC	*Area under ROC the ROC Curve* (Área sob a curva ROC)
CAD	*Computer-Aided Detection* (Detecção Auxiliada por Computador)
CADX	*Computer-Aided Diagnosis* (Diagnóstico Auxiliado por Computador)
CC	Crânio caudal
CNN	*Cellular neural networks* (Redes Neurais Celulares)
DDSM	*Digital Database for Screening Mammography* (Banco de Dados Digital para Análise de Mamografia)
ECM	Exame Clínico das Mamas
FN	Falso Negativo
FNI	Falso Negativo por Imagem
FP	Falso Positivo
FPI	Falso Positivo por Imagem
GNG	*Growing Neural Gas*
INCA	Instituto Nacional do Câncer
MIAS	*Mammography Image Analysis Society* (Sociedade de Análise de Imagem de Mamografia)
MLO	Médio lateral oblíquo
MVS	Máquinas de Vetores de Suporte
MVS-RF	Máquinas de Vetores de Suporte com Feedback Relevância
OMS	Organização Mundial da Saúde
RNAs	Redes Neurais Artificiais
ROC	*Receiver Operating Characteristic* (Característica de Operação do Receptor)
ROI	*Region of interest* (Região de Interesse)
SVFNN	*Support Vector Based Fuzzy Neural Network* (Vetor de Suporte Baseado em Redes Neurais Fuzzy)

VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
WEKA	*Waikato Environment for Knowledge Analysis* (Ambiente para Análise do Conhecimento da Universidade de Waikato)

SUMÁRIO

1 INTRODUÇÃO

No Brasil, as estimativas para o ano de 2009, apontam que ocorrerão aproximadamente 49.000 novos casos de câncer de mama. Este tipo de câncer é provavelmente o mais temido pelas mulheres, devido à sua alta freqüência e os 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, representando nos países ocidentais uma das principais causas de morte em mulheres (INCA, 2008).

As estatísticas indicam o aumento da freqüência do câncer de mama tantos nos países desenvolvidos quanto nos países em desenvolvimento. Segundo a Organização Mundial da Saúde (OMS), nas décadas de 60 e 70 registrou-se um aumento de 10 vezes nas taxas de incidência ajustadas por idade nos registros de câncer de diversos continentes (INCA, 2008).

As formas mais eficazes para detecção precoce do câncer de mama são o exame clínico da mama e a mamografia.

O Exame Clínico das Mamas (ECM) quando realizado por um especialista, pode detectar tumor de até um centímetro, se superficial. Segundo o (INCA, 2009) o ECM deve contemplar os seguintes passos: inspeção estática e dinâmica, palpação das axilas e palpação da mama com a paciente em decúbito dorsal.

A eficiência do exame é proporcional ao grau de habilidade e experiência do profissional para detectar qualquer anormalidade nas mamas examinadas. Ele deve ser realizado periodicamente e o médico indicará a necessidade de mamografia.

A mamografia é a radiografia da mama que permite a detecção precoce do câncer, por ser capaz de mostrar lesões em fase inicial, na faixa de milímetros. É realizada em um aparelho de raios-X apropriado, chamado mamógrafo. Nele, a mama é comprimida de forma a fornecer melhores imagens, e, portanto, melhor capacidade de diagnóstico. Em um exame as duas mamas são analisadas separadamente, e de cada mama são geradas

duas imagens, a primeira é conhecida como visão Crânio Caudal (CC) que é ortogonal ao eixo vertical da paciente, e a segunda é a Médio Laterial Oblíqua (MLO) que é feita em um ângulo oblíquo em relação ao eixo vertical da paciente.

Estudos sobre a efetividade da mamografia sempre utilizam o exame clínico como exame adicional, o que torna difícil distinguir a sensibilidade do método como estratégia isolada de rastreamento (INCA, 2008). No entanto, a sua precisão 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 especialista.

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 fonte de informação 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 Redes Neurais Celulares para segmentar as regiões de interesse, em seguida extrai características de geometria dessas regiões e descreve sua textura através da função K de Ripley, índice de Moran e coeficiente de Geary. Finalmente, o método de aprendizado supervisionado Máquina de Vetores de Suporte (MVS), é utilizado para 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.

A qualidade dos resultados obtidos a partir deste trabalho poderá tornar possível a incorporação da presente metodologia em uma ferramenta para a área médica, servindo como uma ferramenta de apoio ao especialista, principalmente nos casos de difícil visualização.

1.2 Trabalhos Relacionados

A eficiência dos sistemas CAD/CADx são dependentes das técnicas de segmentação e extração de características. A literatura disponível traz trabalhos reconhecidos que tratam do mesmo problema abordado pelo método proposto, ou seja, desenvolver métodos computacionais que possam auxiliar o especialista na tarefa de detecção de lesões em imagens mamográficas.

Redes Neurais Celulares (CNN – Cellular Neural Networks) foram utilizadas em (CHUA e ROSKA, 2004) para testar duas características do câncer. A primeira encontra microcalcificações e a segunda encontra estruturas com forma de espinho ao redor de um dado tumor, indicando a presença de vasos sanguíneos característicos.

Em (SERHAT, ONUR e YILMAZ, 2005), massas foram detectadas usando duas etapas. A primeira etapa procura *pixels* nas imagens em oito direções e regiões de interesse foram identificadas através de limiares. Então, um *template* foi usado para categorizar a região de interesse como massas ou não-massas. Para a análise do desempenho desta técnica foram utilizadas 52 imagens do banco de imagens público MIAS (*Mammography Image Analysis Society*) (SUCKLING, et al. 1994). Essa metodologia obteve uma sensibilidade de 93%, 90% e 81% com taxas de 1.3, 0.7 e 0,33 falsos positivos por imagem respectivamente.

Um algoritmo que combina duas técnicas de extração de características para detecção de massas baseados em intensidade de pixel e textura é apresentado em (TÓTH, TAKÁCS e PATAKI, 2005). O algoritmo analisou 2092 imagens do banco de imagens público DDSM (Digital Database for Screening Mammography) (HEATH, et al. 1998). Este método obteve uma acurácia de 95,1% e uma taxa de 4.3 falsos positivos por imagem.

Em (BELLOTTI. et al. 2006) é apresentado um sistema CAD automático para detecção de massas através de um algoritmo de detecção de bordas para a segmentação, matrizes de co-ocorrência para a descrição e uma rede neural feed-forward de duas camadas treinadas através de gradiente descendente para classificar os candidatos em massa ou não-massa. Foram utilizadas 1093 imagens do banco IMAGIC-5. Na fase de segmentação, 1151 massas e 10377 não-massas foram selecionadas. Este sistema CAD obteve 82% de sensibilidade, uma taxa de 2,8 falsos positivos por imagem e a área sob a curva ROC Az=0,862.

Em (COSTA, BARROS e SILVA, 2007) é comparada a eficiência da MVS e da Análise Discriminante Linear (ADL). Foram utilizadas 200 regiões de interesse (50 malignas, 50 benignas e 100 normais) de imagens mamográficas fornecidas pelos bancos de imagem MIAS e 3600 regiões de interesse (900 malignas, 900 benignas e 1800 normais) extraídas do DDSM. Em todas as bases, metade das regiões de interesse foi utilizada para treinamento e a outra metade para teste. Os resultados da detecção utilizando MIAS alcançaram um desempenho de 85% e 97% para ADL e MVS, respectivamente. Utilizando DDSM, os autores alcançaram 89,2% e 99,6% para ADL e MVS, respectivamente. Com estes resultados os pesquisadores comprovaram o maior desempenho da MVS na detecção de massas nas imagens utilizadas.

Em (ELTONSY, TOURASSI e ELMAGHRABY, 2007) é proposto um algoritmo baseado em múltiplas camadas concêntricas para detecção de massas em mamografias. A técnica é baseada na presença de camadas concêntricas ao redor de uma área focal com características morfológicas suspeitas. Regiões da mama com alta concentração de camadas concêntricas e com queda progressiva da média da intensidade dos pixels são consideradas suspeitas. Esta metodologia utilizou 270 imagens com projeção crauniocaudal do DDSM. Os autores reportaram uma sensibilidade de 81% para detecção e uma taxa de 0,6 falsos positivos por imagem.

Um esquema híbrido utilizado em (HASSANIEN, 2007) utilizou uma combinação de conjuntos fuzzy-rough e estatísticas obtidas através matrizes de co-ocorrência (máxima probabilidade, contraste, momento de diferença inversa, segundo momento angular e entropia) para detectar massas em

mamografias. Esse esquema utilizou 320 imagens do banco de imagens MIAS. Essas imagens foram divididas em 10 grupos, e através do método *Leave-N-Out* (Seção 2.7), regras foram geradas para realizar a detecção com uma acurácia média de 98,46%.

Um algoritmo de detecção de massas é proposto em (KOM, TIEDEU e KOM, 2007) que utiliza um filtro de transformação linear para melhorar a imagem e então usa um limiar local adaptativo na diferença entre a imagem original e a imagem melhorada para realizar a segmentação. A metodologia apresentada utilizou 61 imagens de um banco proprietário, obtendo uma sensibilidade de 95,91%.

Mudanças temporais entre duas mamografias consecutivas são utilizadas para a detecção de massas em (TIMP, VARELA e KARSSEMEIJER, 2007). Foram desenvolvidos dois tipos de características temporais: características diferenciais e características de similiaridade. Características diferenciais indicam a mudança relativa entre duas visualizações. As características de similaridade medem a semelhança em áreas que contêm lesão em duas imagens. Foi utilizado um banco de imagens proprietário, onde os exames foram realizados no período de 1996 a 2000. Foram usados 465 pares (temporais) de mamografias, 238 benignas e 227 malignas. Com o uso de MVS e das características propostas foi possível analisar a mudança temporal e com isso a detecção de massas. Os autores reportaram a área sobre a curva ROC (Receiver Operating Characteristic) com 0.74 sem características temporais e 0.77 com as características temporais.

É proposto em (YING, XINBO e JIE, 2007) um novo esquema para detecção de massas em mamografias utilizando características de intensidade dos *pixels* (contraste, momento invariante, média, gradiente, desvio padrão, momentos e gradiente médio das bordas), geometria (circularidade, densidade; esfericidade; descritor de Fourier) e descritores de textura (*Template* de *Law*, matriz de co-ocorrência, transformada *Wavelet*) para caracterizar regiões suspeitas e MVS-RF (máquina de vetores de suporte com *feedback* de relevância) para a classificação de regiões de interesse em massa ou não-massa. O banco de imagens DDSM foi utilizado e 192 imagens foram usadas

para treinamento e 150 imagens para teste. Este método alcançou 90,6% de sensibilidade e uma taxa de 3,6 falsos positivos por imagem.

Em (IREANEUS e THAMARAI, 2008) é proposto um método de detecção de massas em mamografias utilizando a transformada de wavelet discreta, técnicas de inteligência artificial (análise de dimensões fractais, algoritmo dos “cachorros e coelhos”) e redes neurais artificiais (back propagation). Foram utilizadas 50 imagens do banco de imagens MIAS. Esta metodologia obteve 92% de acurácia na detecção de massas em mamografias.

É apresentada em (NASCIMENTO e RAMOS, 2008) uma comparação do desempenho da detecção de massas através do método Random forest implementado no software WEKA (Waikato Environment for Knowledge Analysis) (WITTEN e FRANK, 2005) utilizando informações extraídas através da transformada de ridgelet (entropia, energia, soma-média, soma-variância e tendência de agrupamento) para as visões crânio caudal e médio lateral oblíqua separadamente e agrupando essas informações. O Banco de imagens utilizado foi o DDSM, onde foram selecionadas 270 regiões de interesse contendo massas e áreas sadias. O desempenho dessa metodologia atingiu 94,4% de sensibilidade, 96,9% de especificidade e 91,8% de acurácia.

Em (MARTINS, et al. 2009) é proposta uma técnica utilizando Growing Neural Gas (GNG) para segmentação de candidatos a massa e Máquinas de Vetores de Suporte (MVS) em conjunto com a função K de Ripley para detectar massas em mamografias. Neste trabalho os autores utilizaram 997 imagens do banco DDSM, onde 436 foram separadas para o treinamento e teste dos classificadores MVS; e 561 foram usadas para avaliar o processo de detecção de massas, obtendo 89,30% de sensibilidade, 0,93 falsos positivos por imagem e 0,02 falsos negativos por imagem.

Em (NUNES, 2009) é proposta uma metodologia de detecção de massas que utiliza o algoritmo de agrupamento K-means e a técnica de Template Matching para segmentar as regiões suspeitas de conterem massas. A metodologia foi testada com 650 imagens mamográficas obtidas da base de dados DDSM. A etapa de segmentação das regiões de interesse conseguiu segmentar 603 massas da amostra, o que equivale a 92,77% dos casos, e

também selecionou 2076 não-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. Finalmente, essas informações são submetidas a uma MVS para que as regiões suspeitas sejam classificadas em massas ou não-massas. A etapa de treinamento e teste foi realizada através de seis diferentes tamanhos dos conjuntos de treino/teste, foram eles: 30/70, 40/60, 50/50, 60/40, 70/30 e 80/20. A etapa de classificação atingiu em média 83,94% de acurácia, 83,24% de sensibilidade, e 84,14% de especificidade, com taxa de 0,55 falsos positivos por imagem e de 0,17 falsos negativos por imagem.

Os trabalhos relacionados acima indicam pesquisas de novas técnicas de detecção de massas em imagens radiológicas são promissoras.

Um fato importante é o uso de banco de imagens públicas, como o DDSM e o MIAS, onde o DDSM ganha destaque pela quantidade de trabalhos onde foi utilizado. Este banco possui uma grande quantidade imagens, de boa qualidade, além de informações extras referentes à imagem e ao exame. Devido a essa boa aceitação, este banco também foi utilizado neste trabalho.

Como Redes Neurais Celulares foram criadas para serem implementadas em circuitos eletrônicos, esta rede neural oferece a possibilidade da etapa de segmentação ser executada de forma mais ágil.

Percebe-se a importância das técnicas de segmentação e extração de características, especialmente na análise forma e 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

Este trabalho está organizado em mais quatro capítulos, descritos resumidamente a seguir.

O Capítulo 2 aborda a fundamentação teórica necessária ao desenvolvimento da metodologia proposta. Apresenta os conceitos e técnicas de processamento de imagens digitais utilizadas, como redes celulares neurais utilizadas para a segmentação de massas, os descritores de geometria e textura extraídas das regiões de interesse e máquinas de vetores de suporte.

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 a classificação das regiões de interesse em massas ou não-massas.

O Capítulo 4 apresenta e discute os resultados obtidos com a metodologia proposta.

O Capítulo 5 apresenta a conclusão sobre o trabalho, mostrando a eficiência dos métodos utilizados e sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica necessária para a compreensão da metodologia utilizada. Aborda-se o câncer de mama, o exame mamográfico, técnicas de processamento de imagens digitais utilizadas, as características geométricas e de textura que descrevem as áreas suspeitas de conterem massas, a técnica de seleção e redução de características, o método de classificação Máquina de Vetores de Suporte e os indicadores de desempenho utilizados para avaliar a metodologia.

2.1 O Câncer de Mama

A formação do câncer de mama inicia-se quando ocorre um erro no processo de divisão celular, produzindo células alteradas. Essas células multiplicam-se de forma desordenada, invadem tecidos adjacentes, sobrevivem fora da mama, geram aglomerados de células tumorais podendo obstruir vasos linfáticos e veias, podem migrar pela corrente sanguínea e interagir com órgãos à distância, processo conhecido como metástases (Figura 1).

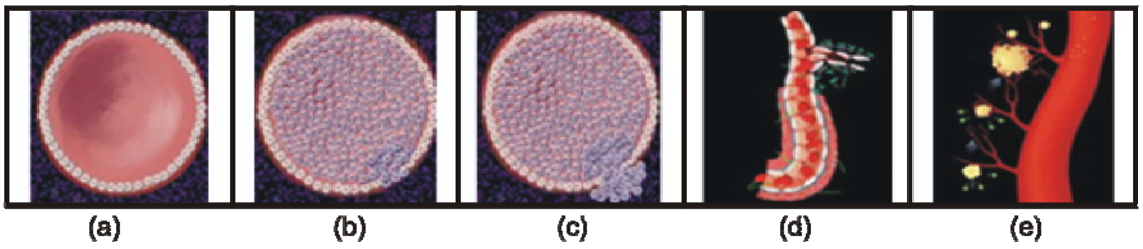


Figura 1: Etapas do desenvolvimento do câncer. (a) Ducto normal; (b) Carcinoma *in situ*; (c) Carcinoma ductal invasor; (d) Invasão de tecidos conectivos; (e) Embolização tumoral no vaso sanguíneo. Fonte: (MAMAINFO, 2009).

Na medida em que as células alteradas invadem um órgão e substituem suas células normais, o órgão começa a ter sua função comprometida.

Segundo a estimativa de câncer no Brasil, para o ano de 2008, surgiram 49.400 novos casos de câncer de mama, um risco estimado de 51 casos a cada 100 mil mulheres. Mundialmente, o câncer de mama é o segundo tipo de câncer mais freqüente no mundo e o mais comum entre as mulheres. A cada ano, cerca de 22% dos casos novos de câncer em mulheres são de mama (INCA, 2009).

O câncer de mama é relativamente raro antes dos 35 anos de idade, mas acima desta faixa etária sua incidência cresce rápida e progressivamente.

Um dos fatores que dificultam o tratamento é o estágio avançado em que a doença é descoberta. Atualmente sabe-se que as chances de cura do câncer de mama são relativamente altas se detectado nos estágios inicias. Estima-se que a sobrevida média geral cumulativa após cinco anos seja de 65% nos países desenvolvidos e de 56% para países em desenvolvimento. Na população mundial, esse índice é de 61% (INCA, 2009).

A detecção precoce é a principal estratégia para controle do câncer de mama. A forma mais eficaz para a detecção precoce do câncer de mama é a mamografia, pois permite que o especialista identifique lesões muito pequenas, em sua fase inicial, da ordem de milímetros (INCA, 2009).

Os tipos de anormalidades observáveis através de uma mamografia podem ser vistos na Figura 2: massas, calcificações e distorções de arquitetura.

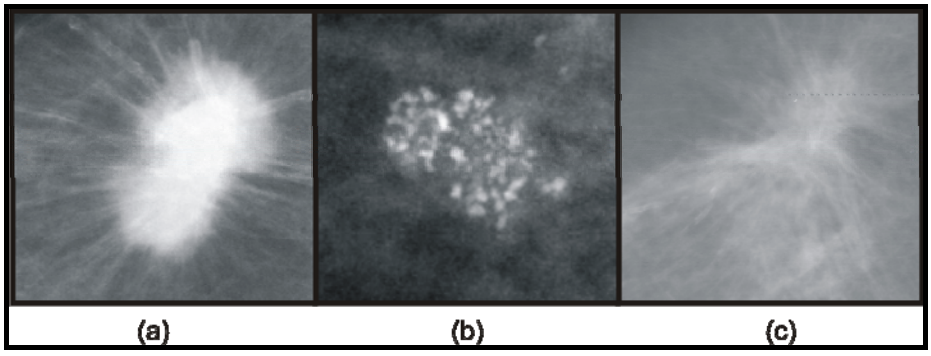


Figura 2: Anormalidades do tecido mamário. (a) Massa espiculada; (b) Microcalcificações; (c) Distorção de arquitetura. Fonte: (DDSM, 2001).

Segundo (PADWAL, 2007), as massas aparecem como regiões densas, de tamanho e formato variáveis, podendo ser classificadas, de acordo com suas bordas em circunscritas, microlobuladas, obscurecidas, mal definidas e espiculadas (Figura 3). Com relação ao formato, podem ser classificadas em redondas, ovais, lobulares ou irregulares. (Figura 4).

Figura 3: Classificação das massas de acordo com o aspecto de suas bordas. Fonte: (PADWAL, 2007).

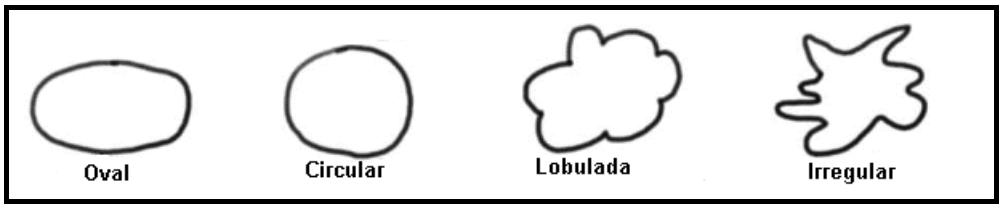


Figura 4: Classificação das massas de acordo com sua forma. Fonte: (PADWAL, 2007).

A lesão benigna é homogênea, de bordas lisas, e empurra o tecido mamário adjacente, sem mudanças secundárias nas mamas. A lesão maligna é uma massa de natureza irregular, que causa alterações secundárias dentro da mama e na pele.

Os sinais secundários podem ser calcificações, espessamento de pele, aumento da vascularização, alterações de aréola e papila ou do estroma mamário, alterações ductais não específicas, demonstrações de nódulos linfáticos axilares e invasão do espaço retromamário. Alguns destes sinais secundários podem estar presentes no exame físico, às vezes muito mais cedo no exame mamográfico, e muitos não são apreciáveis clinicamente. A maioria dos sinais secundários de malignidade são muito úteis; por exemplo, uma

massa de qualquer descrição, com aumento de vascularização, será carcinoma em 75% das vezes (MONTORO, 1979).

Calcificações benignas (Figura 5a) geralmente são maiores, regulares e esparsas. Já as malignas (Figura 5b) são irregulares no tamanho, agrupadas e localizadas em extensa área.

As causas das calcificações malignas são a secreção celular ou a calcificação de células cancerígenas. Dentre as causas da calcificação benigna estão a arterioesclerose, a necrose de gordura, o fibroadenoma involuído e mastite com deposição de cálcio (VALLE,1999).

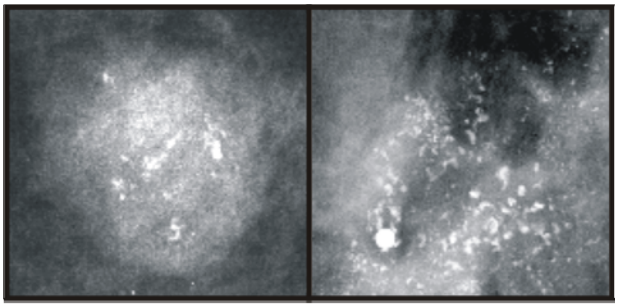


Figura 5: Microcalcificações. (a) Calcificações benignas. (b) Calcificações malignas. Fonte: (DDSM, 2001).

O Quadro 1 contém as principais diferenças entre calcificações malignas e benignas (MONTORO, 1979).

Quadro 1: Características das calcificações mamárias.

Características	**Benignas**	**Malignas**
Densidade	Uniformemente densas	Mais tênues que densas
Bordas	Lisas (Fibroadenoma amorfo)	Pouco nítidas ou irregulares
Disposição espacial	Disseminada, sem padrão definido, podem exibir polaridade	Agrupadas em área restrita, podem ser irregulares, sem polaridade
Número	Poucas e contáveis	Numerosas e incontáveis
Relação com o tumor	Concentradas no centro ou na periferia da lesão	Distribuídas por toda lesão
Localização	Usualmente intraductais	Distribuídas pela massa maligna

2.2 A Mamografia

O objetivo da mamografia é produzir imagens de alta resolução das estruturas internas da mama, a fim de permitir a detecção do câncer de mama. Devido ao fato de que as diferenças de contraste entre tecidos doentes e normais são muito pequenas, esse exame requer um equipamento capaz de realçar tais diferenças e fornecer uma resolução de alto contraste.

A Figura 6 apresenta uma mamografia contendo seus principais elementos constituintes.

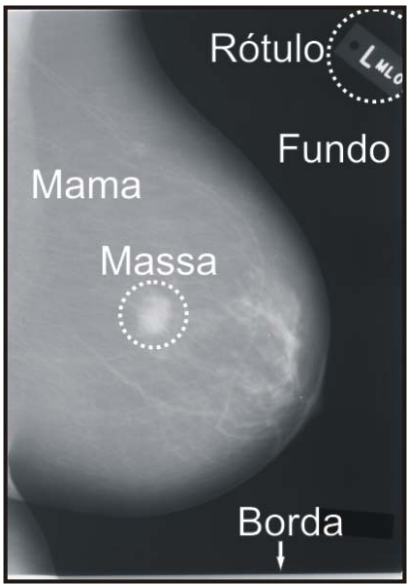


Figura 6: Exemplo de uma mamografia com seus principais elementos. Fonte: adaptado de (DDSM, 2001).

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. Até que sejam planejados sistemas que possam mecanicamente posicionar a mama, são necessários técnicos altamente 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 projeções de cada mama são indispensáveis: uma visão médio-lateral oblíqua (MLO) e uma crânio-caudal (CC), conforme a Figura 7.

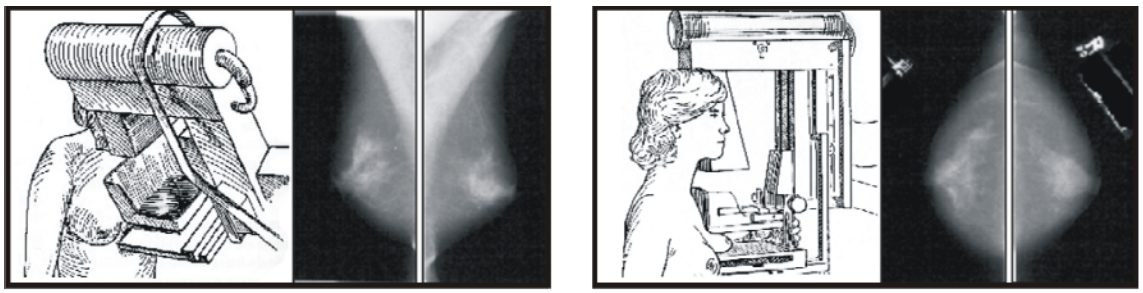


Figura 7: Exames mamográficos. (a) Imagem MLO; (b) Imagem CC. FONTE: adaptado de (ANGELO; 2007).

A projeção MLO é a mais útil, pois permite a visualização do alto da axila para baixo. O termo oblíquo se refere ao plano de compressão da mama. Se a mama não estiver bem posicionada, pode-se deixar de detectar anormalidades. O músculo peitoral deve ser visível, estendendo-se obliquamente até a metade superior da imagem. Além disso, deve ser muito largo no alto e ir se afilando à medida que cruza a parte superior da mama. Estudos sugerem que a mama é representada de maneira ótima quando o músculo peitoral é visível até o eixo do mamilo (KOPANS, 2000).

A projeção CC é a segunda projeção que deve ser obtida rotineiramente. O principal objetivo dessa projeção é obter uma visão da região póstero-medial da mama, complementando a projeção MLO.

A mamografia é um exame de alta sensibilidade. No entanto, sua 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).

Conforme (WANG e KARAYIANNIS, 1998) existe uma grande dificuldade em se trabalhar com imagens de mamografias, as mesmas apresentam baixo contraste. Isto dificulta a interpretação dos resultados por parte do especialista. Estudos (BIRD; WALLACE e YANKASKAS. 1992), (KERLIKOWSKE, 2000) mostram que a mamografia é suscetível a uma alta taxa de falsos positivos como também falsos negativos, causando uma alta proporção de mulheres sem câncer que sofrem uma nova avaliação clínica, enfrentam uma biópsia ou perdem o melhor intervalo de tempo para o tratamento do câncer. 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 de Detecção e Diagnóstico Auxiliado por Computador (CAD / CADx: *Computer-Aided Detection / Diagnosis*), conforme discutida na próxima seção.

2.3 Sistemas CAD/CADx

O desempenho do especialista na interpretação da imagem mamográfica ainda está abaixo do ideal. De fato, é reconhecido um nível de cerca de 10% de diagnósticos falsos-negativos por falha na interpretação da mamografia (NUNES-2. 2001). O problema na análise desse tipo de imagem deve-se principalmente ao baixo contraste das mamografias, à possibilidade de algumas estruturas ficarem “mascaradas” na imagem e à fadiga visual por parte do radiologista (GIGER, 2000). Certamente esse desempenho melhora quando a análise e o diagnóstico em mamografia são elaborados por dois radiologistas (THURFJELL, LENERVALL e TAUBE. 1994) (KARSSEMEIJER, *et al*. 2003), mas este não é um procedimento disponível e possível para todos

os hospitais ou clínicas radiológicas, principalmente devido aos custos e ao tempo gasto nesse tipo de procedimento.

Diagnosticar é muito complexo, pois depende de informações de várias naturezas, tais como experiência médica, indicadores clínicos vindos de imagens, sintomas, laudos patológicos. Muitas vezes o paciente deve se submeter a novos exames complementares desnecessários. Estes exames, além de invasivos, são traumáticos e têm um alto custo financeiro (KARSSEMEIJER, *et al*. 2003).

Sistemas CAD/CADx geralmente incluem técnicas de Processamento de Imagens, Inteligência Artificial e Reconhecimento de Padrões, entre outras. Essas técnicas são aplicadas com o objetivo de melhorar tais imagens e extrair delas informações úteis à detecção de anormalidades e ao diagnóstico (GIGER, 2000).

A Figura 8 apresenta as etapas principais de um sistema de detecção (CAD): obtenção das regiões de interesse (ROI: Region of Interest); caracterização das ROIs; classificação das ROIs através de algum método de reconhecimento de padrões; detecção de massas.

Quando uma área que contém uma possível anormalidade é separada do restante da imagem, o sistema extrai características desse objeto de interesse. 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.

Estima-se que mais de 1500 CADs estejam atualmente em uso em clínicas e hospitais nos EUA para o auxílio no rastreamento do câncer de mama (DOI, 2004). Trabalhos recentes têm mostrado um aumento significativo no desempenho dos radiologistas quando assistidos por um esquema CAD/ CADx.

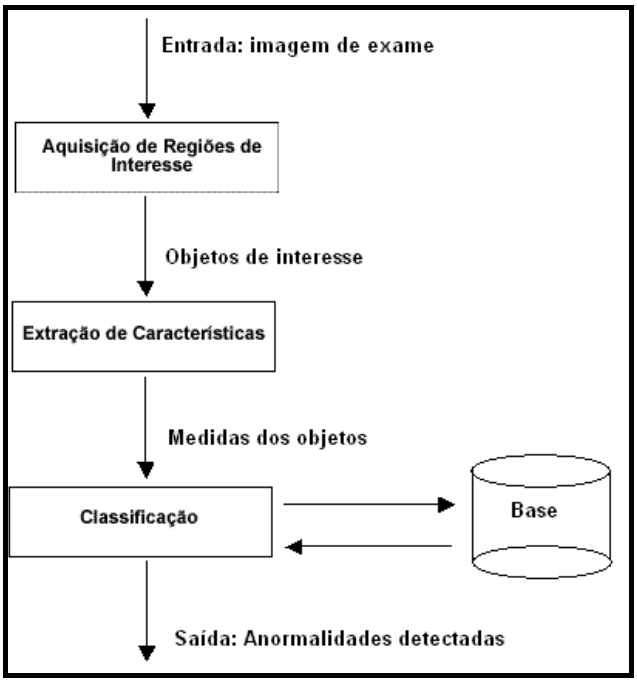


FIGURA 8: Esquema simplificado do funcionamento de um Sistema CAD típico.

Apresenta-se em (FREER e ULISSEY. 2001) a avaliação de diagnósticos em mamografias quando utilizaram um CAD por um período de um ano na rotina clínica. Nesse período, os autores analisaram 12.860 mamografias, seguindo o procedimento de primeiro fornecer o diagnóstico sem o auxílio do CAD e, em seguida, rever o diagnóstico baseado no resultado fornecido pelo CAD. Os resultados da pesquisa mostraram um aumento de 19,5% no número de casos corretamente detectados de câncer de mama quando assistidos pelo CAD, sem um aumento significativo no número de biópsias desnecessariamente realizadas.

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

2.4 Processamento de Imagens

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 processamento de imagens é a atividade de tratar imagens com a finalidade de melhorá-las ou apenas resolver algum problema de modo a permitir que o usuário possa ter acesso a algum tipo de informação que, a princípio, ele não tinha. Didaticamente é dividido em cinco etapas (Figura 9): aquisição, pré-processamento, segmentação, representação e classificação (ALBUQUERQUE E ALBUQUERQUE).

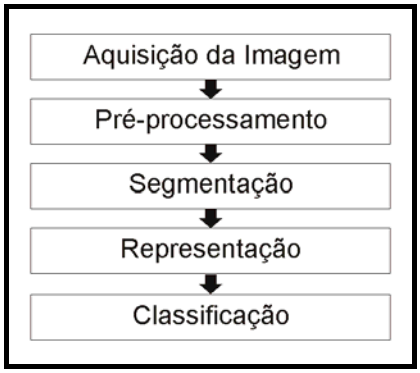


Figura 9: Esquema mostrando as etapas de um sistema de processamento de imagens.

Na aquisição utiliza-se algum mecanismo para gerar as imagens que se deseja processar. As imagens podem ser obtidas tanto através de equipamentos de captura como câmeras, scanners e radares quanto através de simulações por computador. No caso da mamografia, especificamente, os filmes impressos tradicionais podem ser digitalizados por scanners especializados. Neste trabalho 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 pré-processamento tem a finalidade de aumentar a qualidade da imagem, eliminar objetos indesejáveis, tornando mais fácil a sua identificação. Esta etapa é importante porque aumenta a eficiência das etapas posteriores. Nela podemos utilizar diminuição de ruído, realce de contraste, filtros morfológicos, limiarização, dentre outras. Neste trabalho, um procedimento de remoção de ruídos, baseado no algoritmos de agrupamento *K-means* (Seção 2.6.1) e de crescimento de regiões (Seção 2.4.2), foi utilizado para remover os rótulos de identificação do paciente, fundo e bordas do filme que foram digitalizados. Também foram utilizados filtro de Canny (Seção 2.4.3), um operador morfológico de erosão (Seção 2.4.4) e a transformada de Hough (Seção 2.4.5) para a remoção do músculo peitoral em imagens MLO. Além disso, um realce de contraste através equalização do histograma (Seção 2.4.1) foi efetuado para aumentar a discriminação visual das estruturas da mama.

A segmentação permite o isolamento do objeto de estudo, onde os seus resultados são muito importantes na determinação de eventual sucesso ou falha na análise da imagem. Devido à grande quantidade de estruturas diferentes que uma imagem pode ter, a segmentação permite que o processamento seja focado unicamente nas regiões de interesse (*Region of Interest* - ROI). É uma fase delicada do processamento de imagens, pois está intimamente relacionada com características da imagem que são difíceis de traduzir para a máquina. A dificuldade consiste em encontrar medidas consistentes que possam levar a máquina a decidir corretamente a que grupo cada *pixel* pertence. Para esse trabalho utilizamos Redes Neurais Celulares (Seção 2.6.2.3) através de dois *templates* para gerar as regiões suspeitas de conterem massas.

A representação tem a finalidade de extrair da região segmentada um conjunto descritivo de características mensuráveis. Estas características variam muito de acordo com o que se pretende, podendo incluir perímetro, cor dos *pixels*, geometria, etc. Deve-se utilizar medidas 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 geometria (Seção 2.4.7.1) e de textura através da função K de Ripley (Seção 2.4.8), Índices de Moran e Geary (Seção 2.4.8.2).

Por último, na classificação, busca-se através de uma base de conhecimento previamente construída e constituída dos padrões obtidos na etapa de representação, 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 Máquinas de Vetores de Suporte – MVS (Seção 2.6.3) para reconhecer os padrões existentes nas características de geometria e textura das regiões de interesse e classificá-las em massas ou não-massas.

As próximas subseções apresentam técnicas de processamento de imagens digitais utilizadas no desenvolvimento da metodologia proposta neste trabalho.

2.4.1 Realce de contraste

Em uma imagem em tons de cinza, como são as imagens radiológicas, o contraste é uma medida relacionada à distribuição dos tons de cinza dos *pixels*. 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.

Uma técnica bastante utilizada em processamento de imagens digitais é chamada de equalização do histograma.

Segundo (GONZALEZ e WOODS, 2007) o histograma de uma imagem em níveis de cinza na faixa [0; L-1] é uma função discreta $h(r_k) = n_k$, onde r_k é o k-ésimo nível de cinza e n_k é o número de *pixels* na imagem que tem o nível de cinza r_k.

Imagens escuras possuem os componentes do histograma concentrados no lado escuro da escala de cinza (baixa intensidade). Já imagens claras têm os componentes concentrados no lado claro na escala de cinza (alta intensidade). Uma imagem com pouco contraste tem um histograma estreito e concentrado no meio da escala de cinza. Entretanto, uma imagem

com alto contraste apresenta os componentes do histograma cobrindo uma grande faixa na escala e tendo uma distribuição dos *pixels* de maneira aproximadamente uniforme.

Para redistribuir os *pixels* de forma que o histograma da imagem fique equalizado, precisamos utilizar a seguinte equação:

$$S_k = k \times \sum_{j=0}^{k} \frac{n_j}{n} \ , k = 0,1,2,...,L-1 \qquad (1)$$

onde *n* é o número total de *pixels* da imagem e S_k é a nova intensidade do pixel.

Este método de realce é automático e de implementação simples, pois é baseado apenas na informação extraída diretamente da imagem, não necessitando de parâmetros adicionais.

Este trabalho utilizou o método de realce de contraste através da equalização do histograma da imagem para realçar as estruturas internas da mama, melhorando o desempenho das etapas posteriores.

2.4.2 Crescimento de região

O algoritmo de crescimento de regiões é um método de segmentação simples. 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 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 os rótulos de marcação, e na etapa de segmentação, para isolar as regiões de interesse.

2.4.3 Filtro de Canny

O filtro de Canny é um processo de detecção de bordas a partir de critérios de quantificação de desempenho de operadores de bordas conhecidos como os critérios de detecção, localização e resposta múltipla, que corresponde ao fato de que deve haver na saída do operador, uma única resposta para uma única borda. Para que os critérios sejam aproximadamente atendidos, o filtro de Canny aproxima o operador ótimo, obtido a partir dos três critérios de desempenho, pela primeira derivada da função Gaussiana (VALE e POZ , 2002).

Conforme (CANNY, 1986), qualquer filtro para a detecção de bordas deve atender a três critérios básicos. O primeiro deles é denominado Taxa de Erro ou Detecção, consistindo na maximização da razão sinal/ruído (SNR).

O segundo critério especifica que distâncias entre os pontos extraídos pelo detector e as respectivas posições verdadeiras devem ser minimizadas. Tem-se então o critério de Localização (L), definido como sendo o inverso da distância entre um ponto detectado e a respectiva posição verdadeira

Pelo exposto, o projeto de um filtro para a detecção de bordas arbitrárias envolve a maximização de ambos os critérios, o que é equivalente à maximização do produto entre ambos (SNR e L), ficando:

$$\left(\frac{\left| \int_{-w}^{w} G(-x) f(x) dx \right|}{n_0 \sqrt{\int_{-w}^{w} f^2(x) dx}} \right) \cdot \left(\frac{\left| \int_{-w}^{w} G'(-x) f'(x) dx \right|}{n_0 \sqrt{\int_{-w}^{w} f'^2(x) dx}} \right) \quad (2)$$

onde *f(x)* é a resposta de impulso do filtro definido no intervalo *[-w; w], G(x)* é uma borda unidimensional e n_0 a quantificação do ruído da imagem. Assume-se que a borda está centrada em *x* = 0. Na Equação 2, a primeira quantidade entre parêntesis corresponde ao SNR e a segunda à L.

A condição de filtro ótimo (Equação 2) deve ainda atender a um terceiro critério, denominado critério de resposta múltipla. Ou seja, deve haver um único ponto de borda onde exista uma única borda verdadeira. Seja (CANNY, 1986):

$$x_{\max} = 2\pi \left(\frac{\int_{-\infty}^{+\infty} f'^2(x)dx}{\int_{-\infty}^{+\infty} f''^2(x)dx} \right)^{1/2} \quad (3)$$

a expressão matemática para a distância (x_{max}) entre máximos adjacentes na resposta do filtro *f(x)* devido ao ruído. Assim, ao maximizar a condição dada pela Equação 2, deve-se também garantir que x_{max} seja maior possível, aumentando a possibilidade de separação de máximos verdadeiros dos falsos na saída do filtro *f(x)*.

A Figura 10 contém um exemplo de detecção de bordas usando filtro de Canny.

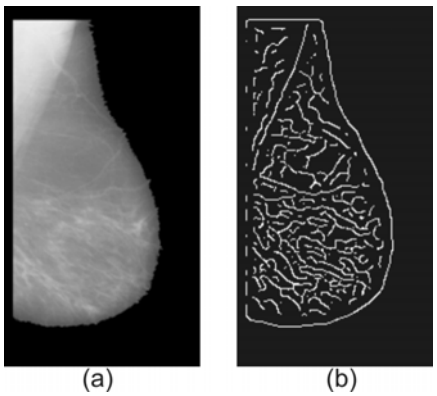


FIGURA 10: Exemplo de detecção de bordas através do filtro de Canny.

Utilizou-se o filtro de Canny neste trabalho para a localização das linhas que possivelmente sejam pertencentes à borda do músculo peitoral em imagens do tipo MLO.

2.4.4 Morfologia matemática

A morfologia matemática ou simplesmente morfologia diz respeito ao ramo de processamento de imagens que se concentra na estrutura geométrica da imagem. Esta estrutura pode ser de natureza macroscópica, onde o intuito é a análise de formas como caracteres impressos, por exemplo, ou pode ser de natureza microscópica onde pode haver interesse na distribuição de partículas ou texturas geradas por pequenas primitivas. Morfologia não é apenas uma teoria matemática, mas uma poderosa técnica de análise de imagens.

Segundo (GONZALEZ e WOODS, 2007) ela é usada predominantemente em: pré-processamento de imagens (filtragem, de ruído, simplificação de formas), segmentação de objetos, detecção da estrutura dos objetos, descrição quantitativa de objetos (área, perímetro).

Para o melhor entendimento de morfologia matemática, é necessário conhecer primeiramente os operadores lógicos que, embora de natureza simples, oferecem um poderoso instrumento em processamento de imagens. Os principais operadores são *AND*, *OR* e *NOT*. Esses operadores podem ser combinados para formar qualquer outro operador lógico, por exemplo, *XOR* e *NOT-AND*. Esses operadores são executados *pixel* a *pixel* entre duas ou mais imagens, exceto para o operador *NOT*, que opera sobre uma única imagem. A Figura 11 contém a aplicação de alguns operadores lógicos em imagens.

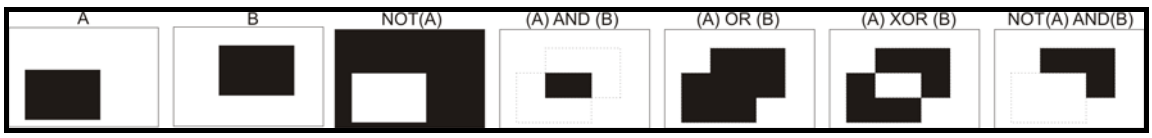


Figura 11: Principais operadores morfológicos. Fonte: adaptado de (GONZALEZ e WOODS, 2007).

Segundo (GONZALEZ e WOODS, 2007) dois operadores extras são bastante utilizados em morfologia matemática: reflexão e translação. A reflexão do conjunto B, denotada $\hat{B}$ é definida como:

$$\hat{B} = \{w \mid w = -b, b \in B\}. \tag{4}$$

A translação do conjunto A pelo ponto z = (z_1,z_2), denotada por $(A)_z$, é definida como:

$$(A)_z = \{c \mid c = a + z, a \in A\}. \tag{5}$$

A base da morfologia matemática consiste em extrair as informações relativas à geometria e a topologia de um conjunto desconhecido (no caso uma imagem) pela transformação através de outro conjunto bem-definido, chamado elemento estruturante.

Existem dois principais operadores morfológicos: erosão e dilatação.

Para os conjuntos *A* e *B* em Z^2 a erosão de *A* por *B* é definida como

$$A \Theta B = \{z \mid (B)_z \subseteq A\} \tag{6}$$

ou seja, a Equação 6 indica que a erosão de *A* por *B* é o conjunto de todos os pontos *z* tal que *B*, transladado por *z*, está contido em *A*. Temos que *A* é a imagem original e *B* o elemento estruturante.

Para os conjuntos A e B em Z^2 a dilatação de A por B é definida como

$$A \oplus B = \{z \mid [(\hat{B})_z \cap A] \subseteq A\} \tag{7}$$

ou seja, a Equação 7 indica que a dilatação de A por B é o conjunto de todos os deslocamentos *z*, tal que a reflexão de *B* intercedida por *A* é sobreposta por pelo menos um elemento. Temos que *A* é a imagem original e *B* o elemento estruturante.

A Figura 12 contém um exemplo dos operadores de erosão e dilatação.

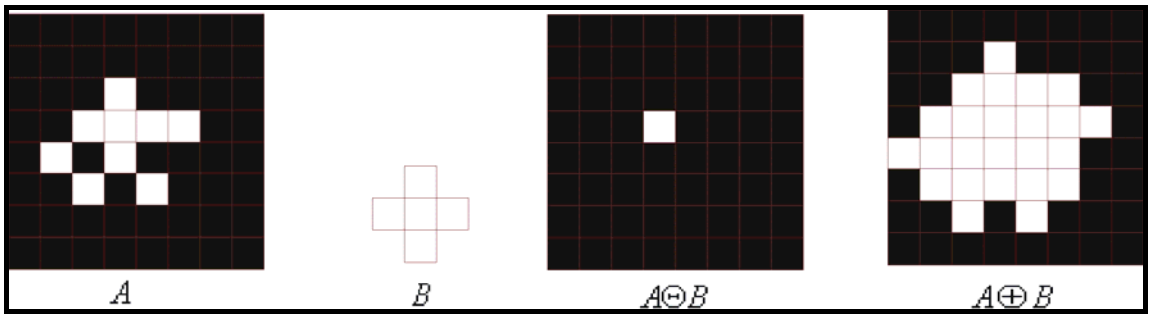


Figura 12: Exemplo da aplicação dos operadores de erosão e dilatação. *A* é a imagem original e *B* é o elemento estruturante.

Utilizou-se o operador de erosão neste trabalho, com o objetivo de se remover todas as possíveis bordas, localizadas pelo filtro de Canny, que não estejam na mesma direção da borda do músculo peitoral.

2.4.5 Transformada de Hough

A transformada de Hough (HOUGH, 1962) é um método padrão para detecção de formas que são facilmente parametrizadas como linhas, círculos, elipses, etc. Em geral, a transformada é aplicada após a imagem sofrer um pré-processamento, comumente a detecção de bordas.

O conceito principal da transformada de Hough está em definir um mapeamento entre o espaço de imagem e o espaço de parâmetros. Cada *pixel* das bordas de uma imagem é transformado pelo mapeamento para determinar células no espaço de parâmetros, indicadas pelas primitivas definidas através do ponto analisado. Essas células são incrementadas e indicarão no final do processo, através da máxima local do acumulador, quais são os parâmetros correspondentes a forma especificada.

Para melhor entendimento da transformada de Hough, citaremos um exemplo prático de detecção de linhas.

Seja a reta r, definida por $y = a.x + b$, como representação paramétrica de uma linha (Figura 13).

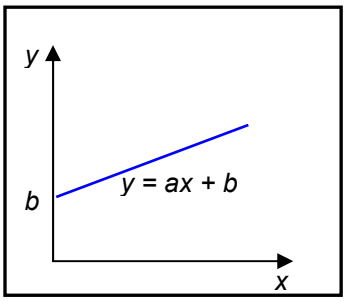


Figura 13: Representação da equação da reta.

A transformada de Hough consiste em mapear um *pixel* da imagem em uma curva no espaço de parâmetros, organizado em forma de um acumulador *n* dimensional, onde *n* corresponde ao número de parâmetros. Como estamos utilizando a reta definida por $y = a.x + b$, então precisamos determinar dois parâmetros: *a* e *b*. Neste caso, o espaço de parâmetros será bidimensional.

Para todos os pontos (*pixels*) da imagem (Figura 14a), no espaço real, calculam-se os parâmetros do espaço de Hough, e acrescenta-se uma unidade na coordenada dos parâmetros da matriz acumuladora *A* (histograma bidimensional). Isso também significa que cada ponto do plano *xy* que foi mapeado, irá corresponder a uma linha no plano *ab* (Figura 14b).

Limitando *a* no intervalo [-2, 3], e empregado os valores de *x* e *y* dos *pixels* da imagem (Figura 14a), calcula-se o valor de *b* através da equação $b = -ax + y$. Os valores são utilizados para incrementar o referido elemento do acumulador *A(a , b)*.

Ao procurar o máximo valor no acumulador *A(a , b)*, verifica-se que este indicará o elemento referenciado por $a = 1$ e $b = 0$, que também indica o local do plano *ab* onde suas linhas contém a maior quantidade de interseções (Figura 14b). Isto significa que a equação $y = 1.x + 0$, ou simplesmente $y = x$, é a equação que melhor define uma reta que passa sobre os pontos *(1,1)* e *(2,2)* (Figura 14a).

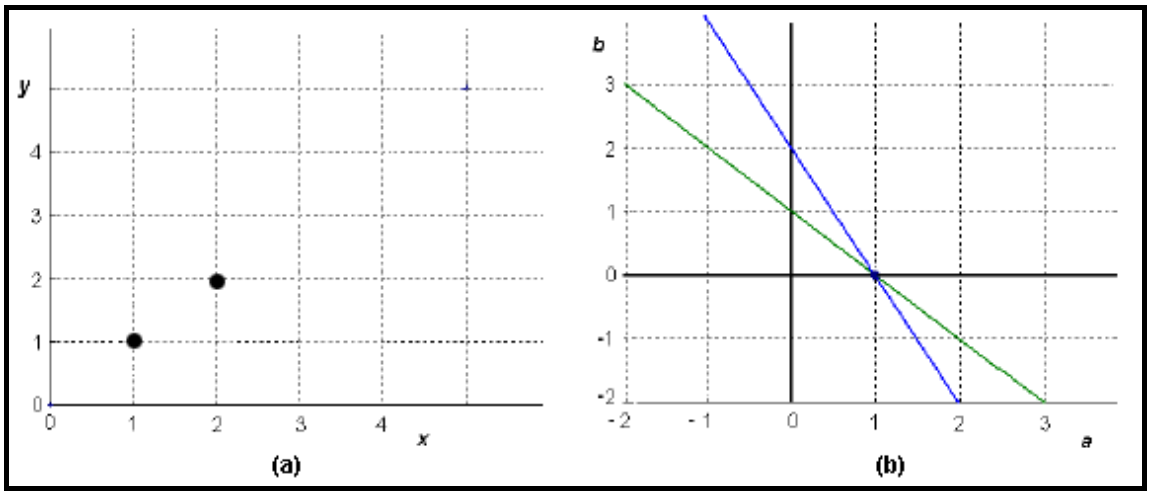


Figura 14: (a) Pontos de uma reta no plano *xy*. (b) Linhas no espaço de Hough (plano *ab*) correspondentes aos pontos do plano *xy*. A linha verde corresponde ao ponto (1,1) e a linha azul ao ponto (2,2).

Uma forma de detecção de linhas sugeridas por (DUDA e HART, 1972), utiliza coordenadas polares para representar linhas da seguinte forma:

$$r = x\cos(q) + y\sin(q) \tag{8}$$

Essa é uma forma mais eficiente de representar uma reta do que a forma $y = a.x + b$, pois necessitamos do ângulo q, que varia num espaço finito de 0 a 180º, para calcular a distância r, diferentemente da forma $y = a.x + b$ onde a e b são parâmetros com valores variando entre $-\infty$ e $+\infty$.

A transformada de Hough foi utilizada neste trabalho para a localização e remoção do músculo peitoral em imagens MLO.

2.4.6 Quantização e amostragem espacial

Uma imagem analógica para tomar o formato digital deve sofrer uma discretização espacial (amostragem) e em amplitude (quantização). Para tanto é feita uma amostragem (normalmente uniforme) da imagem nas direções *x* e *y*, gerando uma matriz de *M x N* pontos seguida de uma quantização em *L* níveis de cinza.

O número de bits por *pixel* corresponde ao número de cores ou tons de cinza que podem ser representados. A quantidade de cores é determinada pela fórmula $L = 2^b$, onde *L* é o número de níveis possíveis e *b* o número de bits da imagem.

Não existem critérios absolutos que nos permitam decidir o número ótimo de *pixels* e bits para amostrar uma determinada imagem. De forma geral, utilizar uma maior resolução aumenta o número de *pixels* para representar uma área e mais informações podem ser extraídas, porém, eleva-se o custo computacional para armazenar a imagem e também o tempo de processamento necessário para a realização de uma análise.

A Figura 15 contém um exemplo de uma imagem em 8, 4 e 2 bits por *pixel*.

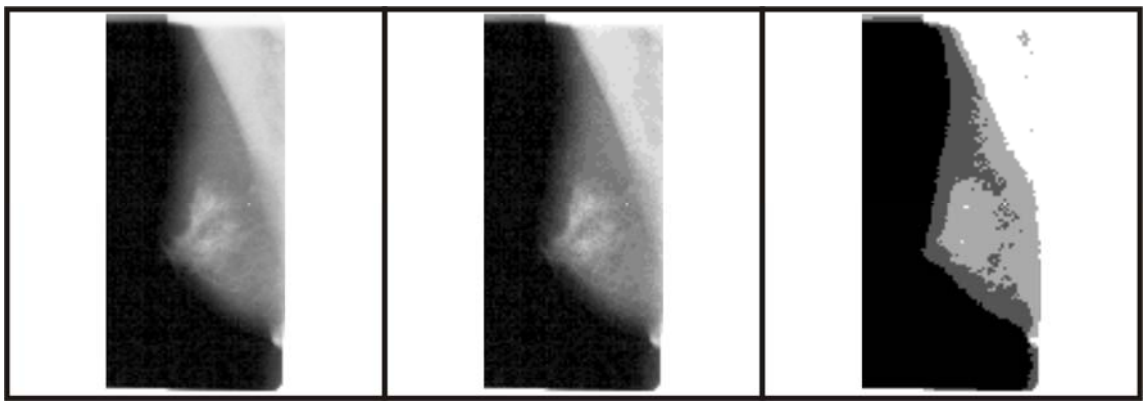


Figura 15: Efeito da quantização numa imagem. Na seqüência imagem com 8, 4 e 2 bits.

Este trabalho utiliza diferentes níveis de quantização das imagens de mamografias para obter uma maior quantidade de relacionamentos de textura na fase de descrição.

2.4.7 Representação e descrição de objetos

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.

As próximas subseções descrevem as características utilizadas neste trabalho.

2.4.7.1 Descritores geométricos

Uma forma de descrever e analisar objetos é através de sua forma. Na literatura de processamento de imagens podem ser encontrados vários descritores ou medidas geométricas (WIRTH, 2001). Entre eles, podem ser destacados compacidade, excentricidade, circularidade, descritores Fourier, descritores baseados em momentos e curvatura (RANGAYYAN, *et al*. 1997).

Para o cálculo de tais medidas não são levados em conta os níveis de cinza presentes nos objetos, ou seja, o objeto é binarizado. Somente suas propriedades geométricas, tais como área e perímetro são utilizados nos cálculos.

Nas subseções seguintes iremos apresentar os descritores de geometria utilizados nesse trabalho com o objetivo de caracterizar massas e regiões sadias localizadas na etapa anterior. São eles: excentricidade, circularidade, compacidade, desproporção circular e densidade circular.

2.4.7.1.1 Excentricidade

Excentricidade é a razão entre o menor e maior eixo horizontal ou vertical, caracterizando como o objeto está distribuído espacialmente entre seus eixos (BRAZ, 2006).

A excentricidade pode ser calculada por

$$E = \frac{(\mu_{02} - \mu_{20}) + 4\mu_{11}}{A} \quad (9)$$

sendo *A* a área do objeto.

Os momentos centrais μ_{pq} são obtidos através de

$$\mu_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (x - \overline{x})^p (y - \overline{y})^q \quad (10)$$

com *p+q >1* e $(\overline{x}, \overline{y})$ sendo o centro de gravidade do objeto em estudo. *M* e *N* correspondem à largura e a altura do objeto.

A Figura 16 compara dois objetos, um com alta e outro com baixa excentricidade.

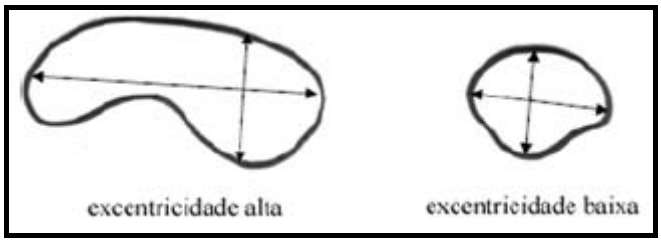


Figura 16: Comparação de objetos através da excentricidade. Fonte: (BRAZ, Júnior. 2006).

2.4.7.2 Circularidade

A circularidade é a medida geométrica que define o grau de proximidade do objeto com um círculo. Trata-se da razão entre a área do objeto e o perímetro convexo, onde o perímetro convexo (Figura 17) é calculado sobre a região que engloba o objeto de forma a não deixar picos ou defeitos no contorno.

A circularidade é definida por

$$C = \frac{4\pi A}{(p_{convexo})^2} \quad (11)$$

sendo *A* 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 o objeto, mais próximo de 1 é o valor de sua circularidade (BRAZ, 2006).

Figura 17: Exemplo de perímetro convexo. (a) Objeto com picos e defeitos. (b) Objeto com superfície convexa. Fonte: (FONSECA, 2001).

A Figura 18 exemplifica o conceito de circularidade através da indicação dos valores dessa medida para dois objetos.

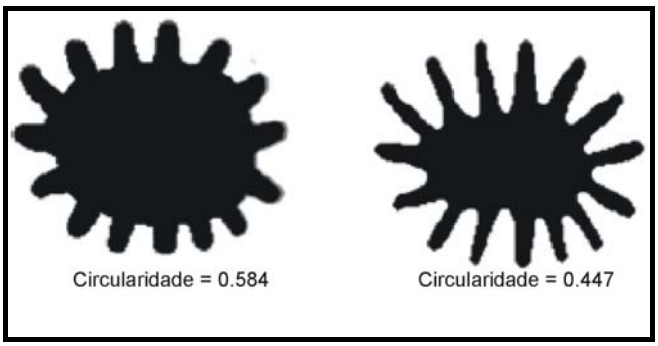


Figura 18: Comparação de objetos através da circularidade. Fonte: (BRAZ, 2006).

2.4.7.3 Compacidade

Compacidade é a medida geométrica que mede a densidade do objeto, em comparação com uma figura perfeitamente densa, ou seja, um círculo (SCHOUTEN, 2003).

A compacidade é definida por

$$C_o = \frac{p^2}{4\pi A} \tag{12}$$

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

A Figura 19 ilustra graficamente o conceito de compacidade.

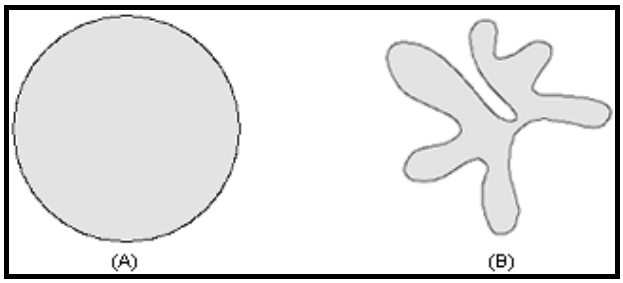


Figura 19: Comparação de objetos através da compacidade: (a) compacto; (b) não-compacto. Fonte: (SCHOUTEN, 2003).

2.4.7.4 Desproporção circular

A desproporção circular (SOUSA, SILVA e PAIVA, 2007), informa o quanto determinado objeto é desproporcional em relação a uma superfície totalmente circular. Ela é obtida através da Equação 13, onde p é o perímetro do objeto em estudo e R_e o raio estimado (Equação 14) do círculo de mesma área do objeto.

$$D = \frac{p}{2\pi R_e} \tag{13}$$

$$R_e = \sqrt{\frac{A}{\pi}} \tag{14}$$

2.4.7.5 Densidade circular

A densidade circular (SOUSA, SILVA e PAIVA, 2007) utiliza um círculo de mesma área e mesmo centro de massa do objeto para estimar qual a porcentagem da interseção entre o objeto e o círculo. Tal medida é calculada através da seguinte fórmula:

$$D_c = \frac{n}{A} \tag{15}$$

onde *A* é a área do objeto e, *n* é o total de pontos pertencentes ao objeto e também ao círculo de raio estimado R_e (Equação 14), onde o círculo e o objeto

têm o mesmo centro de massa. A Figura 20 ilustra o conceito de densidade circular para diferentes objetos.

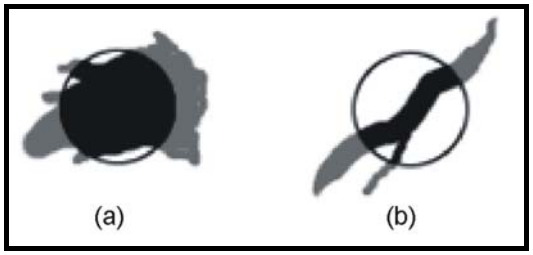


Figura 20: Ilustração da densidade circular. (a) Objeto com alta densidade circular. (b) Objeto com baixa densidade circular. Fonte: (BRAZ, 2006).

A densidade circular assume valores próximos a zero para objetos alongados e valores próximos a 1 para objetos mais circulares.

2.4.8 Descritores de textura

Outra forma de se extrair informações de objetos é através da análise da textura do mesmo.

A textura constitui uma característica diretamente relacionada com as propriedades físicas que a superfície de um objeto apresenta. Ela descreve o padrão de variação de tons de cinza ou cor numa determinada área. Trata-se de um termo intuitivo e de largo emprego, mas que, apesar de sua grande importância, não possui uma definição precisa (EBERT, 1994). 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 textura se caracteriza pela repetição de um modelo sobre uma região, sendo este modelo repetido em sua forma exata ou com pequenas variações. Isso torna a textura um excelente descritor regional, contribuindo para uma melhor precisão dos processos de reconhecimento, descrição e classificação de imagens. Apesar de seus benefícios, seu processo de reconhecimento exige um alto nível de sofisticação e complexidade

computacional (EBERT, 1994). Diversas abordagens para análise ou segmentação, encontradas na literatura, operam sobre texturas. Tais métodos, em geral, são baseados na análise de espectro (ANGELO e HAERTEL, 2001), análise estatística dos *pixels* de uma região (LI, 1995), wavelets (MANDAL, IDRIS e PANCHANATHAN, 1999) e dimensão fractal (CHAUDHURI e SARKAR, 1995).

Para a análise de textura utilizada neste trabalho, foram empregadas as seguintes técnicas de análise espacial: a função K de Ripley, e os índices de correlação espacial de Moran e de Geary.

2.4.8.1 Função K de Ripley

A análise de dados espaciais envolve a exploração, relacionamento, explicação e tendência de padrões sistemáticos, como regularidade, agrupamentos ou aleatoriedade (PAIVA, RODRÍGUEZ e CORREIA, 1999).

A análise de dados espaciais pode ser empreendida sempre que as informações estiverem espacialmente localizadas e quando for preciso levar em conta, explicitamente, a importância do arranjo espacial dos fenômenos na análise ou na interpretação de resultados desejados.

Assim, a análise de padrões de pontos espaciais é uma importante ferramenta para examinar detalhadamente a distribuição de pontos discretos, como por exemplo, *pixels* em uma imagem mapeados para coordenadas cartesianas *(x, y)* em uma região de interesse.

A função K de Ripley é um método de análise de segunda ordem comumente utilizada em análise de dados espaciais. Ela é uma estatística comumente utilizada em ecologia, para descrever a distribuição espacial de árvores e outras espécies em uma floresta. Nos últimos trinta anos, usa aplicação foi utilizada nas mais diversas áreas como, por exemplo, geologia, epidemiologia, geomorfologia, criminologia (LANCASTER e DOWNES, 2004).

Essa função pode ser utilizada para resumir um padrão de pontos, testar hipóteses sobre o padrão, estimar parâmetros e ajustar modelos (RIPLEY, 1977).

A fórmula da função K de Ripley é definida da seguinte forma:

$$K(r) = \frac{A}{n^2} \sum_i \sum_j \delta(d_{ijr}) \quad (16)$$

onde r é o raio de análise, *i* e *j* são pontos distintos ($i \neq j$) pertencentes a uma área *A* da amostra, *n* o total de pontos da amostra e $\delta(d_{ijr})$ uma função que devolve 1 se a distância d_{ij} entre os pontos *i* e *j* é menor do que o raio *r* e 0 em caso contrário. Ou seja, a função K de Ripley conta o número de ocorrências do evento *j* em um círculo de raio *r* para cada centro *i*, conforme visto na Figura 21.

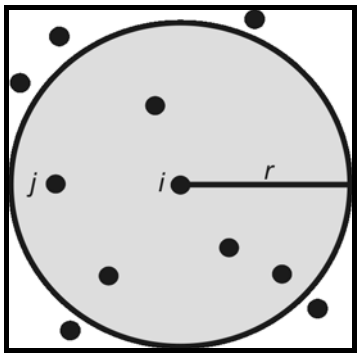


Figura 21: Ilustração da análise através da função K de Ripley para um raio r dado. Fonte: (MARTINS, *et al*. 2009).

Como cada ponto da região é tomado uma vez para ser o centro do círculo, a função *K*(r) provê uma inferência em nível global sobre a área em estudo. Entretanto, esta medida também pode ser considerada em uma forma local para o i-ésimo ponto, conforme a Equação 17. Dessa maneira, é possível descrever a textura de uma região em uma imagem através da função local K de Ripley. A partir da escolha de um centro *i*, são examinados as ocorrências de *pixels* de um mesmo nível de cinza *j*, para diferentes valores de raios *r*.

$$K_i(r) = \frac{A}{n} \sum_{i \neq j} \delta(d_{ijr}) \quad (17)$$

Cada nível de cinza (padrão espacial) é examinado separadamente dos demais, e tratado como a ocorrência ou não de um evento dentro da distância *r* especificada. Assim, o número de elementos do vetor de características obtido através do uso de *K(r)* é dado pelo número de níveis de cinza presentes na imagem vezes o número de raios desejado.

O presente trabalho utiliza a modificação sobre a função *K(r)*, através da análise dos padrões de pontos em anéis, ao invés de círculos, conforme proposto em (MARTINS, *et al*. 2009). A modificação consiste em substituir a região de interesse da Equação 17 pela região compreendida entre dois círculos concêntricos (Figura 22). Tal modificação mostrou-se superior à função *K(r)* tradicional na caracterização de tecidos da mama para a classificação em massa e não-massa.

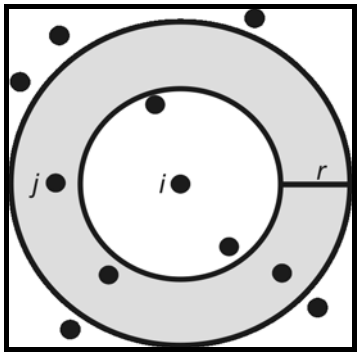


Figura 22: Função K de Ripley modificada para um dado raio *r*. Fonte: (MARTINS, *et al*. 2009).

Segundo (MARTINS, *et al*; 2009) a superioridade da modificação pode ser explicada pelo fato de que o uso de círculos como regiões de estudo acaba fornecendo informação cumulativa para regiões periféricas do objeto. Em contrapartida, o uso de anéis concêntricos elimina possíveis interferências de regiões centrais, fornecendo informação mais precisa a respeito das regiões periféricas do objeto.

2.4.8.2 Índices de Moran e Geary

A grande maioria dos fenômenos não se distribui no espaço casualmente. Por exemplo, o relevo de uma região pode ser resultado de falhas geológicas; a presença de uma espécie vegetal pode depender de certas combinações de clima e solo; a concentração de população num país pode refletir o roteiro histórico de sua colonização; a industrialização em determinados estados pode demonstrar a existência de recursos naturais. Relevo, espécies vegetais, população e indústrias apresentam-se, portanto, organizados de uma maneira que pode ser prevista, pelo menos em parte, uma vez conhecidos os seus determinantes (ANSELIN; 1995).

Uma das técnicas mais utilizadas no estudo de fenômenos espaciais é a Análise de Autocorrelação Espacial. A autocorrelação espacial refere-se à redundância de informação entre duas realizações de um fenômeno quando elas ocorrem próximas uma da outra. Sua presença distorce os resultados obtidos pela aplicação de modelos estatísticos tradicionais baseados na hipótese de independência entre as realizações da variável de interesse (CÂMARA, *et al*. 2004).

Esta técnica permite identificar a estrutura de correlação espacial que melhor descreve o padrão de distribuição dos dados. A idéia básica é estimar a magnitude da Autocorrelação Espacial entre as áreas, evidenciando como os valores estão correlacionados no espaço. Ou seja, estimar quanto do valor observado de um atributo numa região é dependente dos valores dessa mesma variável, nas localizações vizinhas. Enquadram-se nesta categoria o Índice Global Moran e o Índice de Geary.

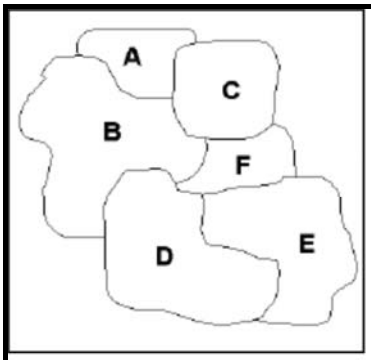
O índice de Moran (*I*) é a estatística mais difundida e mede a autocorrelação espacial a partir do produto dos desvios das variáveis de interesse em relação à média. Há outras medidas que apesar de similares, medem a dependência espacial a partir de outras operações, diferença simples como no índice de Geary ou soma simples como nas estatísticas *G* (ANSELIN; 1995).

Formalmente, *I* é escrito como:

$$I = \frac{n}{W}\left(\frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2}\right) \text{ para } i \neq j \qquad (18)$$

onde *n* é o número de observações, w_{ij} é o elemento na matriz de proximidade para o par *i* e *j*, *W* é a soma dos ponderadores da matriz de proximidade, z_i e z_j são os desvios em relação à média, ou seja, $z_i = x_i - \overline{x}$ e $z_j = x_j - \overline{x}$.

A informação espacial é incorporada no modelo a partir da matriz de proximidade *W*. A proximidade pode ser definida de diferentes maneiras: distância euclidiana, tempo de viagem ou acessibilidade. A mais comumente utilizada define proximidade a partir da propriedade topológica de contigüidade (Figura 23). *W* é uma matriz binária, onde 1 está associado às zonas com fronteiras em comum e 0 àquelas sem esta propriedade. Como a matriz de proximidade é utilizada em cálculos de indicadores de análise exploratória, por conveniência, ela é muitas vezes utilizada normalizada por linha, ou seja, com a soma dos ponderadores de cada linha igual a 1 (CÂMARA, *et al*. 2004).



(a)

	A	B	C	D	E	F
A	0	1	1	0	0	0
B	1	0	1	1	0	1
C	1	1	0	0	0	1
D	0	1	0	0	1	1
E	0	0	0	1	0	1
F	0	1	1	1	1	0

(b)

	A	B	C	D	E	F
A	0	0,5	0,5	0	0	0
B	0,25	0	0,25	0,25	0	0,25
C	0,33	0,33	0	0	0	0,34
D	0	0,33	0	0	0,33	0,34
E	0	0	0	0,5	0	0,5
F	0	0,25	0,25	0,25	0,25	0

(c)

Figura 23: Cálculo da matriz de proximidade. (a) mapa com vizinhança; (b) matriz de proximidade binarizada; (c) Matriz de proximidade normalizada.

De uma forma geral, o índice de Moran presta-se a um teste cuja hipótese nula é de independência espacial (aleatoriedade); neste caso, seu valor seria zero. Valores positivos [0;1] indicam correlação direta, ou seja, valores de uma variável em áreas próximas tendem a serem semelhantes. Já

para valores negativos [-1;0] temos correlação inversa. Ela indica que valores de uma variável em áreas próximas tendem a serem diferentes.

O Índice de Geary permite identificar a presença da influência das variáveis associadas a uma localização e mensurar a intensidade dessa influência sobre as realizações dessas mesmas variáveis em localizações próximas.

O Índice de Geary é dado pela Equação 19 (GRIFFITH, 1987):

$$G = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{2(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij})\sum_{i=1}^{n} z_i^{\,2}} \text{ para i}\neq\text{j} \qquad (19)$$

onde x_i e x_j são os valores da variável de interesse nas áreas *i* e *j*, x é a média da variável de interesse em todas as áreas da amostra, *n* é o número de áreas na amostra, w_{ij} é o pertencente a matriz de proximidade *W* e $z_i = x_i - \overline{x}$

O índice de Geary apresenta caso de independência espacial para valores iguais a 1 (aleatoriedade). Valores menores que um [0;1] indicam correlação direta, ou seja, valores de uma variável em áreas próximas tendem a serem semelhantes. Já para valores maiores que um [1;2] temos correlação inversa, que indica que valores de uma variável em áreas próximas tendem a serem diferentes.

2.5 Seleção de Características

A qualidade dos resultados de um determinado classificador depende da relevância das características consideradas no conjunto de exemplos de treinamento *T*. Em problemas de classificação, o conjunto dados *T* é composto por *m* instâncias (*x,y*), onde *x* é um vetor contendo as informação extraídas dos objetos em estudo, e *y* é a classe a qual *x* pertence.

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).

Porém, essa abordagem apresenta alguns problemas. Conforme (LANGLEY e IBA, 1993), quanto maior a quantidade de características irrelevantes, maior a necessidade de exemplos de treinamento para alcançar uma dada acurácia. Deve-se observar que alguns dados são raros ou de difícil obtenção ou mesmo de custo oneroso. Outro problema está com relação ao desempenho, maior quantidade de informação, maior a necessidade de equipamentos mais sofisticados e um maior tempo de processamento, o que pode ser inviável financeiramente ou em sistemas de tempo real. Também é comum que a maioria das características disponíveis não seja informativa o suficiente para a distinção entre as diferentes classes (XING, 2003). Isto ocorre quando as informações são irrelevantes, redundantes ou apresentarem níveis elevados de ruído.

Para solucionar esses problemas, em reconhecimento de padrões, costuma-se realizar uma seleção das características mais relevantes, a fim de aumentar a eficiência do classificador e diminuir os custos de processamento. (EFROYMSON, 1960). Este trabalho utiliza o algoritmo de seleção de variáveis chamado de *stepwise.*

2.5.1 Algoritmo *Stepwise*

O algoritmo de seleção de características *stepwise* é apropriado quando se tem um elevado número de variáveis candidatas a variáveis explicativas, daí a escolha de tal método neste estudo. Esse método indica o conjunto mais provável de variáveis explicativas.

Em cada etapa, é removida uma única variável. O critério de remoção de uma variável é usualmente adotado em termos de teste *F*-parcial (WERKEMA e AGUIAR. 1996). O método é finalizado quando não há mais variáveis a serem excluídas do conjunto. O processo é controlado pela escolha dos parâmetros *F* (valores estatísticos), que seguem a distribuição *F* de Snedecor, para a remoção de uma variável do modelo, ou o que é equivalente,

por meio da determinação dos níveis de significância, α_s, associados aos valores de *F.*

O método tem início com o ajuste de um modelo de regressão linear simples para cada uma das *P*-1 variáveis explanatórias X_k. Para cada modelo a estatística *F** é calculada:

$$F_k^* = \frac{MS_M(X_k)}{MS_R(X_k)} \tag{20}$$

MS_M e MS_R são os erros médios quadráticos do modelo e residual, respectivamente. A variável X_k com o maior valor de *F** é candidata para a primeira adição. Se este valor de *F** ultrapassar determinado valor F_{IN}, então, a variável é adicionada no modelo. Caso contrário, o programa é concluído e não são incluídas variáveis no modelo. Se este valor de *F** for menor do que um determinado valor F_{OUT}, a variável X_k é removida do modelo, caso contrário, ela permanece.

De forma geral F_{IN} pertence ao intervalo [2,4]. Já F_{OUT} é pertencente ao intervalo $0.0 < F_{OUT} \leq F_{IN}$, onde o valor 2 é recomendado tanto para F_{IN} quanto para F_{OUT}. Esses valores são informados pelo usuário no início da execução do método.

2.6 Reconhecimento de padrões

Reconhecimento de padrões é um sub-tópico da aprendizagem de máquina cujo objetivo é classificar informações, ou padrões, baseado ou em conhecimento *a priori* ou em informações estatísticas extraídas dos padrões. Essa área de atuação é estudada por vários campos, tais como psicologia, etologia e ciência da computação.

Um sistema completo de reconhecimento de padrões consiste de um sensor que obtém observações a serem classificadas ou descritas; um mecanismo de extração de características que computa informações numéricas ou simbólicas das observações; e um esquema de classificação das observações, que depende das características extraídas.

O esquema de classificação é geralmente baseado na disponibilidade de um conjunto de padrões que foram anteriormente classificados, chamado de conjunto de treinamento, neste caso temos um aprendizado supervisionado. O aprendizado pode também ser não supervisionado, de forma que o sistema não recebe informações *a priori* dos padrões, estabelecendo então as classes dos padrões através de análise de padrões estatísticos.

2.6.1 K-means

O K-means é uma técnica de agrupamento de dados não supervisionada que consiste no particionamento de um conjunto de dados em k-subconjuntos, ou grupos, de forma que os dados dentro de cada grupo compartilhem características semelhantes em relação a alguma medida de proximidade. É uma técnica comumente utilizada para análise estatística de dados, sendo útil em diferentes áreas, incluindo aprendizado de máquina, mineração de dados, reconhecimento de padrões, análise de imagens e bioinformática (JAIN, MURTY e FLYNN, 1999).

O algoritmo K-means, baseado na técnica estatística do centróide, toma um parâmetro inicial *k* e divide um conjunto de dados em *k* grupos de tal forma que a similaridade dentro dos grupos seja alta enquanto que a similaridade entre os grupos seja pequena (SOUKUP e DAVIDSON, 2002). O procedimento do algoritmo é, em primeiro lugar, selecionar aleatoriamente *k* objetos do conjunto de dados (Figura 24a), onde cada um destes objetos representará inicialmente uma média do centro de cada grupo. A seguir os demais objetos são classificados nos grupos para os quais apresentam maior similaridade (HAN e KAMBER, 2006). Um parâmetro tipicamente utilizado como medida de similaridade é a distância Euclidiana, podendo-se também utilizar-se a distância Manhattan. Assim o agrupamento dos dados se dá através das menores distâncias em relação ao centro dos grupos, ou seja, cada elemento é classificado no grupo para o qual apresente a menor distância ao seu centro (Figura 24b). O processo interativo consiste na atualização dos centros de grupos (Figura 24c), através da média das coordenadas dos pontos

de cada grupo, até a função objetivo seja atingida, indicando que os centros pararam de se mover (Figura 24d).

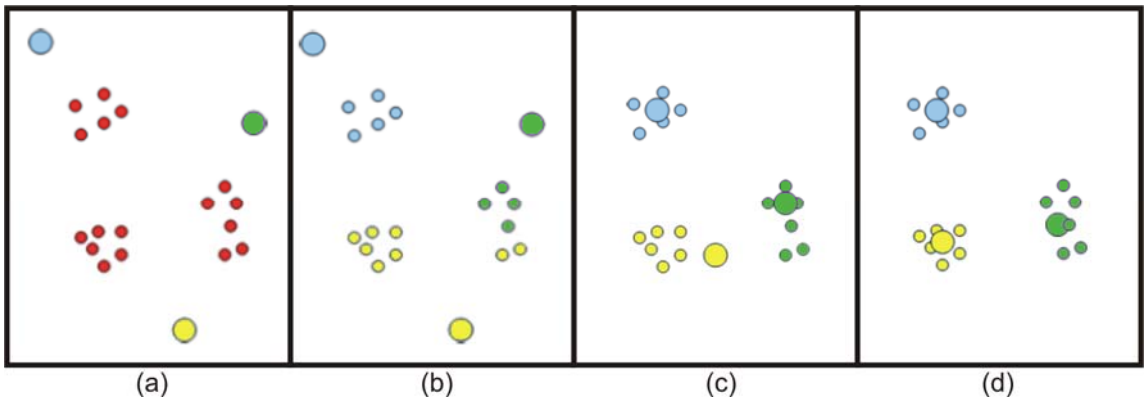


Figura 24: Esquema ilustrativo da execução do K-means. (a) Dados originais (vermelho) e os centros criados aleatoriamente (azul, verde e amarelo). (b) Dados atribuídos aos centros mais próximos. (c) Movimentação dos centros e reatribuição dos dados. (d) Fim do processo com os centros estacionados.

A função objetivo definida pelo K-means é o método dos mínimos quadrados:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2 \qquad (21)$$

onde $\| x_i^{(j)} - c_j \|^2$ é a medida de distância entre um ponto qualquer $x_i^{(j)}$ até o centro do grupo c_j em questão; *k* é a quantidade de grupos. Então a função objetivo *J* indica a distância dos *n* objetos contidos em seus respectivos grupos. Quanto menor o valor da função objetivo, melhor o agrupamento.

A próxima seção irá detalhar as Redes neurais celulares como uma ferramenta de processamento de imagens. Para isso será abordada uma breve revisão sobre redes neurais artificiais.

2.6.2 Redes neurais

O estudo do cérebro é atrativo, sob o ponto de vista da computação, por propiciar o desenvolvimento de modelos de processamento de informação biologicamente inspirados, como é o caso das Redes Neurais Artificiais (RNAs). As RNAs representam uma tentativa de superar limitações que o computador digital apresenta, buscando, para isso, imitar os princípios de funcionamento do cérebro. As principais características do cérebro são: alto grau de paralelismo; operação mesmo sob sinais corrompidos por ruído; robustez e tolerância a falhas, graças à redundância e operação descentralizada dos neurônios; capacidade de adaptação e auto-organização, conseguidos com a experiência e/ou aprendizado. De um ponto de vista computacional, o cérebro é um computador analógico que, ao contrário de computadores digitais que processam símbolos, utiliza sinais de origem eletro-química para efetuar computações.

O estudo de RNAs se utiliza de modelos simplificados, mas adequados para fins de investigação da capacidade computacional do cérebro, principalmente quanto à resolução de problemas ou aplicações reais.

2.6.2.1 O neurônio biológico

O comportamento de um neurônio depende do estado em que ele se encontra e do grau de estimulação que recebe. Estímulos vindo de outros neurônios chegam como um sinal eletro-químico até a membrana da célula nervosa, através dos dendritos. Se estes estímulos atingirem um determinado limiar (*threshold*), a célula dispara um potencial de ação, que são descargas elétricas geradas no prolongamento da célula denominado axônio. Estes axônios, que se encaminham até a outra extremidade do neurônio, se ramificam para estabelecer comunicação com células vizinhas. Tais ramificações, que se encontram muito próximas, não estão exatamente em contato com as outras células, e esse espaço que as separam é denominado de sinapse. Nas sinapses, o estímulo elétrico que percorre o neurônio, se transforma em um estímulo químico, por meio da liberação de substâncias

neurotransmissoras pelo neurônio. Estas substâncias, por sua vez, atuam na estimulação dos dendritos de outras células nervosas as quais o neurônio considerado possui sinapses, propagando, dessa forma, o sinal recebido (CORRÊA, 2004).

Um modelo de neurônio utilizado em RNAs, entretanto, abstrai todos esses detalhes de como se procede o fenômeno biológico, como os detalhes da descarga de potenciais de ação, o período refratário (período em que o neurônio não dispara, após ser estimulado), os atrasos na transmissão de sinais intra-neuronais, entre outros.

2.6.2.2 O neurônio artificial

De maneira simplificada, um neurônio artificial é um elemento que efetua uma soma ponderada de sinais e que produz uma saída, de acordo com uma função de ativação. Embora um neurônio real emita pulsos de potencial de ação (sinal contínuo), no modelo considerado, esta informação é simplificada, sendo que o estado do neurônio é considerado binário, indicando apenas se ele está disparando potenciais de ação ou não. Esse disparo é controlado pela função de ativação sinal (Figura 25b). Experimentalmente, observa-se que existe uma relação entre a quantidade média de estímulos que um neurônio recebe e a quantidade média de potenciais de ação que ele produz, sendo que essa relação toma uma forma sigmoidal (Figura 25a). Outras funções de ativação, que são normalmente utilizadas em modelos de neurônio, também possuem uma forma semelhante a esta (Figura 25).

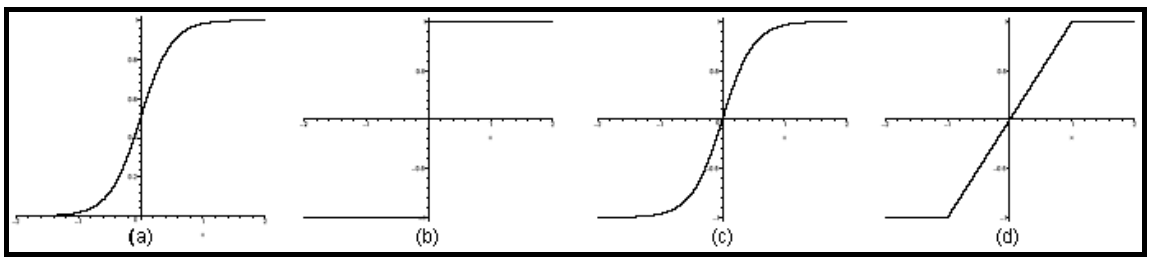


Figura 25: Funções de ativação. (a) sigmoidal; (b) sinal; (c) hiperbólica; (d) saturação. Fonte: (CORRÊA, 2004).

Matematicamente, este o modelo é expresso pela seguinte equação:

$$x_i(n+1) = f(\sum_j^n w_{ij} x_j(n) - \theta_i)\text{, } i = 1, \ldots, n \qquad (22)$$

onde $x_i \in B = \{-1,1\}$ é o estado do i-ésimo neurônio (i = 1, ..., n), disparando o potencial de ação (x_i = 1) ou em repouso (x_i = -1). A variável n indica o tempo (discreto). A variável contínua w_{ij} indica o peso (eficiência) da conexão entre os neurônios i e j. θ indica o limiar da operação do neurônio i.

As RNAs constituem um paradigma de computação diferente do paradigma convencional, que se baseia em um elemento processador central (CPU) controlando todo o sistema. No paradigma neural, o processamento é realizado de forma distribuída, por meio de neurônios artificiais. Diferentemente de um fluxo seqüencial de instruções que um computador digital tem que seguir, nas RNAs têm-se elementos processadores que operam de maneira paralela, interagindo uns com os outros. Não existe um programa explícito, sendo que o funcionamento do modelo depende da dinâmica dos neurônios, e da forma como esses são conectados, o que determina o tipo de tarefa que a rede realizará. Normalmente, ao invés de ser programada, uma rede neural aprende a resolver uma tarefa que lhe é atribuída por meio de um algoritmo de treinamento ou aprendizado, que ajusta os parâmetros da mesma.

Os modelos de RNAs podem ser divididos em duas classes: os dinâmicos e os estáticos. Nos modelos estáticos (*Feedforward*) a rede é vista como uma função, onde sua operação é estabelecer uma relação estímulo-resposta de acordo com a entrada da rede. Nesse caso, os elementos processadores são conectados em uma só direção (Figura 26a). Nos modelos dinâmicos a rede é vista como um sistema dinâmico, isto é, sua operação não é apenas dada em função da entrada, mas também do estado em que a rede se encontra (Figura 26b). Dessa forma, diz-se que tais redes possuem memória.

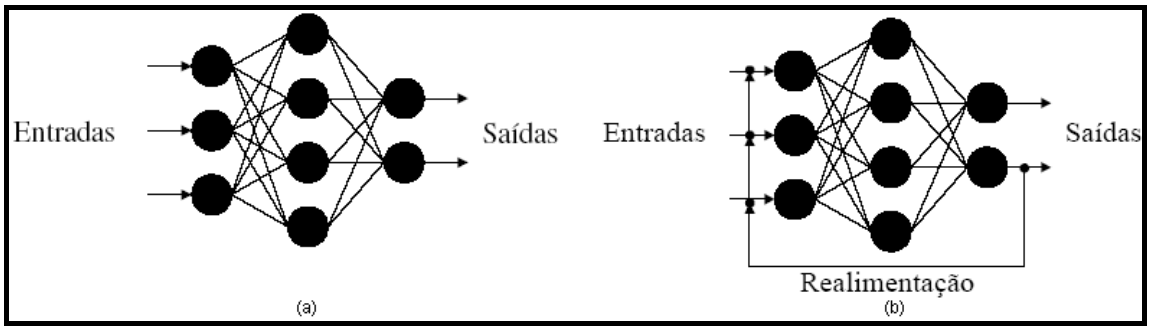


Figura 26: Modelos de redes neurais. (a) Modelo estático; (b) Modelo dinâmico. Fonte: (CORRÊA, 2004).

As redes neurais celulares fazem parte dos modelos dinâmicos e serão apresentadas na subseção a seguir.

2.6.2.3 Redes Neurais Celulares

Para se explorar todo o paralelismo de uma RNA, sua implementação deve ser feita em *hardware*. No entanto, isso normalmente é inviável, devido ao alto número de conexões que podem existir entre os elementos processadores. No modelo de Hopfield, por exemplo, os neurônios são totalmente conectados uns aos outros, produzindo, para uma rede com n neurônios, um total de n^2 conexões. Como o crescimento do número de conexões, em função do número de neurônios, é de ordem quadrática ($O(n^2)$), a construção de uma rede de Hopfield torna-se inviável, para n grande. As alternativas propostas para resolver tal problema consistem basicamente de: técnicas de multiplexação de sinais (*hardwares* ópticos, por exemplo); diferentes arquiteturas/topologias de interconexões, ambas visando à diminuição do número de conexões entre os elementos processadores. Redes neurais celulares exploram essa segunda alternativa de solução ao problema discutido.

Em 1988, Leon Chua e Lin Yang propuseram o modelo de Redes Neurais Celulares (do inglês Cellular Neural Network - CNN) (CHUA e YANG, 1988), que se tratam de redes localmente acopladas, isto é, redes cujas interconexões entre os neurônios se restringem a uma certa vizinhança dos mesmos. Esse tipo de rede é ilustrado na Figura 27, onde neurônios são

representados por círculos preenchidos, enquanto as conexões entre elas são representadas por arestas ligando tais círculos.

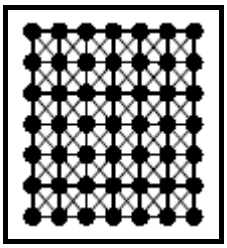


Figura 27: Ilustração de uma CNN. Fonte: (CORRÊA, 2004).

As Redes Neurais Celulares são assim denominadas por apresentarem certa semelhança com os Autômatos Celulares (sistemas dinâmicos discretos no tempo e espaço, localmente acoplados, cujos estados também são discretos), em relação à conectividade local presente entre as células da rede. Essa característica, em conjunto com uma implementação em *hardware* analógico da rede, fazem com que as CNNs sejam especialmente eficientes na execução de aplicações que envolvam grandes quantidades de cálculos realizados apenas à nível local, ou seja, envolvendo apenas dados de pontos vizinhos. Esse é o caso, por exemplo, de aplicações em Processamento de Imagens.

Segundo (CHUA e ROSKA, 2004) a arquitetura padrão de uma rede neural celular consiste de uma matriz retangular $M \times N$ de células $C(i, j)$ com coordenadas cartesianas (i, j), tendo $i = 1, 2, ..., M$ e $j = 1, 2, ..., N$. Cada célula é definida matematicamente por:

$$\begin{cases} dx_{ij}(t) = -x_{ij}(t) + \sum_{C(k,l)\in S_r(i,j)} A_{i,j;k,l} y_{kl}(t) + \sum_{C(k,l)\in S_r(i,j)} B_{i,j;k,l} u_{kl} + z_{ij} \\ y_{ij} = \frac{1}{2}\left|x_{ij}+1\right| - \frac{1}{2}\left|x_{ij}-1\right| \end{cases} \quad (23)$$

Onde t é o tempo (passo de interação), $S_r(i, j)$ é esfera de influência de raio r $(r \in \mathbb{R}^{+}_{*})$ da célula $C(i, j)$ definida como sendo o conjunto de todas as células

vizinhas *(2r+1)* × *(2r+1)* com centro em *(i, j)*. Exemplos de vizinhança entre células são mostradas na Figura 28.

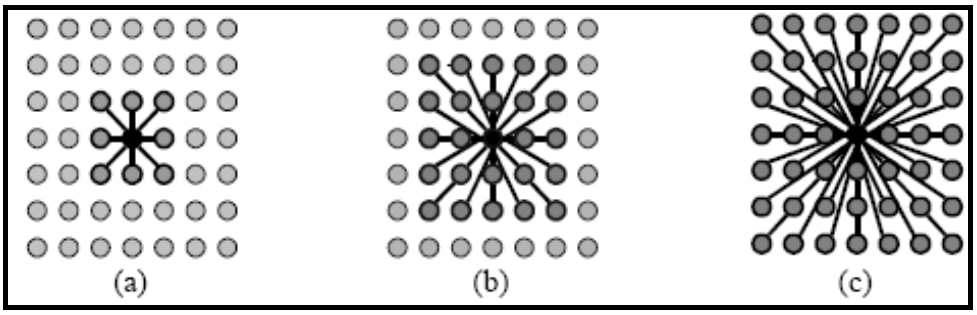


Figura 28: Exemplos de vizinhanças de uma célula para diferentes valores do raio *r*. (a) *r* = 1; (b) *r* = 2; (c) *r* = 3; Fonte: (CORRÊA, 2004).

As variáveis x_{ij}, y_{kl}, u_{kl} e $z \in \square$ são chamadas de *estado*, *saída*, *entrada* e *limiar* da célula *C(i, j)* respectivamente. $A_{i,j;k,l}$ e $B_{i,j;k,l}$ são chamados operadores de *feedback* e *entrada sináptica*.

A relação entre x_{ij} e y_{ij} é mostrada na Figura 29, também chamada de não-linearidade padrão.

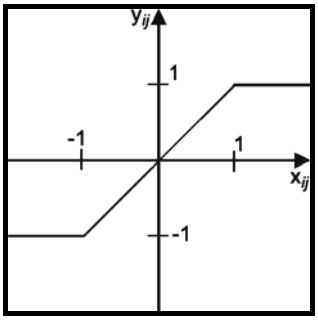


Figura 29: Gráfico da função Não-linearidade padrão. Fonte: (CHUA e ROSKA, Tamás; 2004).

A entrada u_{kl} é a intensidade do *pixel* de uma imagem *M* × *N* em escala de cinza, normalizado para a faixa de $-1 \leq u_{kl} \leq +1$, onde o branco é codificado como -1 e o preto como +1.

A equação 23 irá repetir-se até não corra variação de estado em qualquer célula *C(i, j),* entre os tempos *t* e *t-1.*

A Figura 30 contém uma ilustração do cálculo do estado x_{ij}, em um determinado tempo *t* para $S_1(i, j)$.

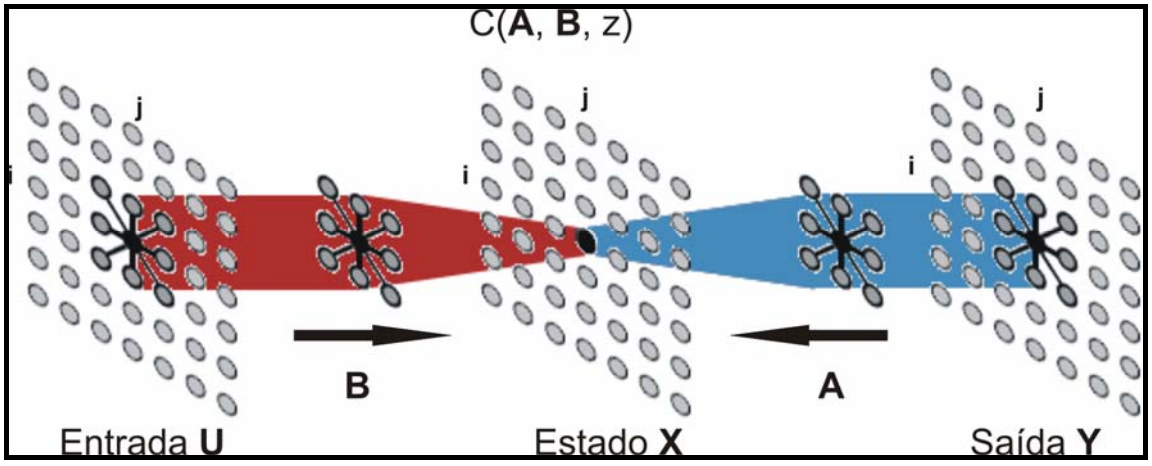


Figura 30: Ilustração do cálculo do estado x_{ij} qualquer. Fonte: adaptado de (CORRÊA, 2004).

Segundo (CHUA e ROSKA, 2004) os primeiros anos da introdução do paradigma CNN, alguns *templates* foram desenvolvidos usando conhecimento prévio, tentativa-erro ou usando simuladores para calcular CNNs dinâmicas. Hoje alguns métodos estão disponíveis para gerar *templates*. Dentre eles podemos citar métodos simbólicos para imagens binárias utilizando tabelas verdade, algoritmos genéticos, modelos neuromórficos baseados em organismos vivos, técnicas fuzzy, técnicas de redes neurais, etc.

Ilustraremos a seguir o desenvolvimento, através de conhecimento prévio, de um *template* para remoção de ruídos em imagens, baseado no operador Laplaciano.

Sejam os operadores *A* e *B* de raio *r*=1 representados por:

$$A = \begin{pmatrix} A_{-1,-1} & A_{-1,0} & A_{-1,1} \\ A_{0,-1} & A_{0,0} & A_{0,1} \\ A_{1,-1} & A_{1,0} & A_{1,1} \end{pmatrix} \text{ e } B = \begin{pmatrix} B_{-1,-1} & B_{-1,0} & B_{-1,1} \\ B_{0,-1} & B_{0,0} & B_{0,1} \\ B_{1,-1} & B_{1,0} & B_{1,1} \end{pmatrix} \quad (24)$$

Os elementos centrais ($A_{0,0}$ e $B_{0,0}$) denotam a conexão de cada célula consigo mesma.

O Laplaciano de uma função L(x,y) definida no plano é dado pela soma das derivadas parciais segundas de L (Equação 25).

$$\Delta^2 L = \frac{\partial^2 L}{\partial x^2} + \frac{\partial^2 L}{\partial y^2} \tag{25}$$

Discretizando esse operador através da expansão da série de Taylor, temos:

$$\Delta^2 L \approx \frac{1}{\delta h^2}\left[L(x+\delta h, y) + L(x, y+\delta h) + L(x-\delta h, y) + L(x, y-\delta h) - 4L(x, y)\right] \tag{26}$$

onde δh é o passo de discretização. Assim o operador A ficará da seguinte forma:

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix} \tag{27}$$

Os outros parâmetros da rede (Equação 27) são B = 0, z = 0.

Este trabalho utiliza Redes Neurais Celulares para localizar áreas em imagens mamográficas que potencialmente podem conter massas.

2.6.3 Máquina de vetores de suporte - MVS

A Máquina de Vetores de Suporte – MVS (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.

Seus algoritmos de treinamento 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, 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.

A Figura 31 mostra hiperplanos de separação entre duas classes linearmente separáveis. O hiperplano ótimo (linha central) separa as duas classes e mantém a maior distância possível com relação aos pontos da amostra.

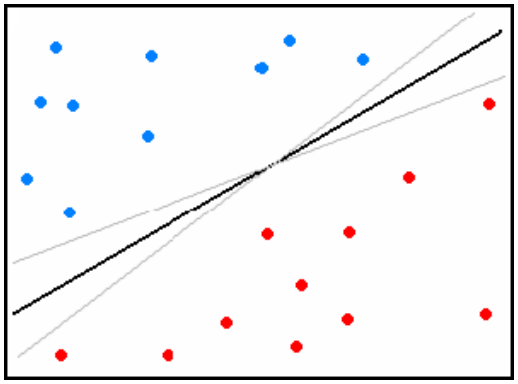


Figura 31: Separação de duas classes através de hiperplanos.

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 28 seja satisfeita:

$$y_i((w.x_i)+b) \geq 1 \tag{28}$$

$$\Phi(w) = \frac{w^2}{2} \tag{29}$$

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 29, de acordo com a restrição dada pela Equação 28 (CHAVES, 2006).

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 28 sejam quebradas.

O problema de otimização passa a ser então a minimização da Equação 30, de acordo com a restrição imposta pela Equação 28, 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.

$$\Phi(w,\xi) = \frac{w^2}{2} + C\sum_{i=1}^{N}\xi_i \tag{30}$$

$$y_i((w.x_i)+b)+\xi \geq 1 \tag{31}$$

onde *C* é uma penalidade para a função Φ, ξ_i é a variável de folga que suaviza as restrições dada pela Equação 31 , e *N* é o número de amostras de entrada.

Através da teoria dos multiplicadores de *Lagrange*, chega-se à Equação 32. O objetivo então passa a ser encontrar os multiplicadores de *Lagrange* α_i ótimos que satisfaçam a Equação 33 (CHAVES, 2006):

$$w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i, x_j) \quad (32)$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \ , \ 0 \leq \alpha_i \leq C \quad (33)$$

Apenas os pontos onde a restrição da Equação 28 seja exatamente iguais a 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 32 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.

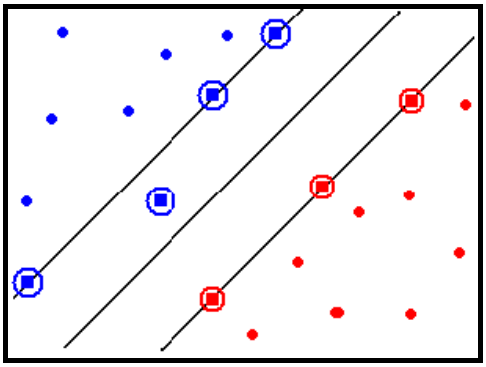


Figura 32: Vetores de suporte para determinação do hiperplano de separação.

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, 2008). A função *K* pode realizar o mapeamento das

amostras para um espaço de dimensão muito elevada sem aumentar a complexidade dos cálculos. A Equação 34 mostra o resultado da Equação 32 com a utilização de um núcleo *K*.

$$w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (34)$$

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 35:

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \quad (35)$$

2.7 Validação cruzada

Após um classificador ter sido projetado e configurado é necessário saber até que ponto pode-se confiar nos seus resultados. Para isto, procura-se estimar qual o seu erro em condições normais de funcionamento e não nas condições de treino.

Então um ponto crucial para análise do desempenho do classificador é saber como dividir a amostra em treino e teste. Existem vários métodos de partição da amostra com a finalidade de avaliação, também conhecidas como técnicas de validação cruzada. Cada uma contém suas características, vantagens e desvantagens. Entre as mais conhecidas temos: re-substituição, *hold-out*, *leave-N-out*, *bootstrap*, *leave-one-out* (KIRALJ e FERREIRA, 2009).

Este trabalho utilizou a técnica do *leave-N-out* para avaliação da metodologia utilizada.

Com o método *leave-N-out* o conjunto de dados é dividido igualmente em *N* subconjuntos, e o treino efetua-se concatenando *N*-1 subconjuntos, e a validação usando o subconjunto restante. As fases de treino e teste são depois repetidas *N* vezes, permutando-se circularmente os subconjuntos. O erro final é

calculado usando a média dos erros de cada fase. Usa-se geralmente *N*=10. Esta técnica proporciona uma avaliação menos tendenciosa do erro do classificador à custa de maiores custos computacionais.

2.8 Métricas de desempenho

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*), média de falsos positivos por imagem (*FPI*) e média de falsos negativos por imagem (*FNI*).

2.8.1 Acurácia

A acurácia mede a porcentagem total de casos corretamente classificados (Equação 36).

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{36}$$

2.8.2 Sensibilidade

A sensibilidade mede o desempenho da classificação em relação aos casos positivos (Equação 37).

$$S = \frac{VP}{VP + FN} \tag{37}$$

2.8.3 Especificidade

A especificidade mede o desempenho da classificação em relação aos casos negativos (Equação 38).

$$E = \frac{VN}{VN + FP} \tag{38}$$

2.8.4 Falsos positivos por imagem e Falsos negativos por imagem (FPI e FNI)

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).

$$FPI = \frac{\sum_{i=1}^{n} i_{FP}}{n} \tag{39}$$

$$FNI = \frac{\sum_{i=1}^{n} i_{FN}}{n} \tag{40}$$

onde *i* é a i-ésima imagem analisada e *n* o número total de imagens. *FP* e *FN* é a quantidade de falsos positivos e falsos negativos, respectivamente, da imagem *i*.

2.8.5 Overlay

Outra medida utilizada para avaliar a taxa do acerto sobre as áreas localizadas é chamada de *overlay* (do inglês encobrir). Ela indica a proporção média do tamanho das áreas localizadas em relação às áreas reais. Se ela for maior que 1 indica que em média, as áreas localizadas são maiores que as áreas originais. Se ela for menor que 1 indica que, em média, as áreas localizadas são menores que as áreas originais. A Equação 41 define a medida *overlay O*.

$$O = \frac{\sum_{i=1}^{n} \frac{Ns_i}{Nr_i}}{n} \tag{41}$$

onde Nr_i é o número de pontos da área marcada pelo especialista na imagem *i*, e Ns_i é o número de pontos da área selecionada na imagem *i*, e *n* é o número de imagens que tiveram as massas encontradas.

2.8.6 Curva ROC

Em sistemas de apoio a decisão, principalmente nas áreas de Ciências Médicas e Ciências da Saúde, a avaliação de desempenho de sistemas classificadores é realizada com a análise ROC (MAZUROWSKI *et al*, 2008), que é uma análise mais robusta, relacionando a sensibilidade e a especificidade do classificador.

A análise da curva ROC (do inglês *Receiver Operating Characteristic*) oferece ferramentas para selecionar possíveis modelos ótimos e descartar modelos sub-ótimos independentemente do custo do contexto ou distribuição da classe. É uma forma natural de analisar o custo/benefício na tomada de decisões em diagnósticos. Ela foi primeiramente desenvolvida por engenheiros elétricos e engenheiros de radares durante a 2ª guerra mundial para detecção de objetos inimigos nos campos de batalha, também conhecida como teoria da detecção de sinais.

Na teoria de detecção de sinais, a curva ROC é um gráfico da sensibilidade X (1-especificidade) para sistemas com classificadores binários quando o limite de descriminação é variado.

Um modelo de classificação (ou diagnóstico) é um mapeamento de instâncias em uma determinada classe (ou grupo). O resultado da classificação pode ser um valor real (saída contínua) na qual o limite entre classes deve ser determinado por um valor de limiar, por exemplo, determinar se uma pessoa tem hipertensão baseado em medida de pressão sanguínea; ou pode ser um rótulo discreto que indica uma classe.

Considerando o problema de predição em duas classes (classificação binária) no qual os resultados ou estão rotulados como classe positiva (*p*) ou negativa (*n*). Há quatro possíveis resultados de um classificador binário: *VP*, *FP*, *VN* e *FN* (ver Seção 2.8).

Duas distribuições hipotéticas de padrões são apresentadas na Figura 33.

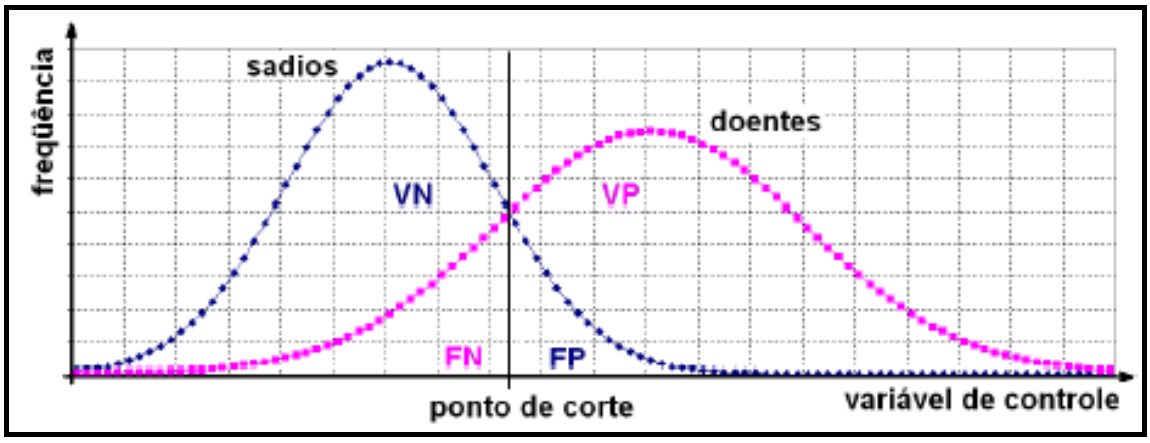


Figura 33: Exemplo de duas distribuições, doentes e sadios, em função de uma variável de controle e com um ponto de corte para a classificação. Fonte: (SOVIERZOSKI, ARGOUD e AZEVEDO, 2008).

O classificador utiliza o ponto de corte (*pc*) para efetuar a discriminação entre as classes. Padrões com valor acima do *pc* são classificados como positivos e abaixo do *pc* são considerados negativos. Os

padrões da distribuição dos doentes com valor acima do *pc* são classificados como verdadeiros positivos (*VP*) e abaixo do *pc* são classificados como falsos negativos (*FN*). Os padrões da distribuição dos sadios com valor abaixo do *pc* são classificados como verdadeiros negativos (*VN*), e acima do *pc* são classificados como falsos positivos (*FP*). As classificações corretas (acertos do classificador) são *VP* e *VN*. As classificações erradas (erros do classificador) são FP e FN (SOARES, REZENDE e FORTES, 2008).

A matriz de confusão com as indicações do especialista humano e as indicações do exame (MEDRONHO *et al*. 2008) é apresentada na Quadro 2, permitindo quantificar as distribuições da Figura 33.

Através da matriz de confusão cada padrão é classificado e totalizado numa das quatro categorias (*VP, VN, FP* e *FN*), compondo os indicadores estatísticos de desempenho do classificador, em função do valor do *pc* utilizado.

Quadro 2: Matriz de confusão.

		Especialista	
		Doente	Sadio
Indicação do exame	Positivo (doente)	*VP*	*FP*
	Negativo (sadio)	*FN*	*VN*

Através da matriz de confusão cada padrão é classificado e totalizado numa das quatro categorias (*VP, VN, FP* e *FN*), compondo os indicadores estatísticos de desempenho do classificador, em função do valor do *pc* utilizado.

Para cada valor do *pc*, na Figura 33, existe um valor de sensibilidade e especificidade correspondente, calculado pelas Equações 37 e 38 respectivamente. Para pequenos valores do *pc* (Figura 34, o valor de *FN* é baixo e *FP* é elevado, resultando em alta sensibilidade e baixa especificidade.

Aumentando-se gradativamente o valor do *pc* ocorre a inversão de comportamento dos índices estatísticos, como apresenta a Figura 35.

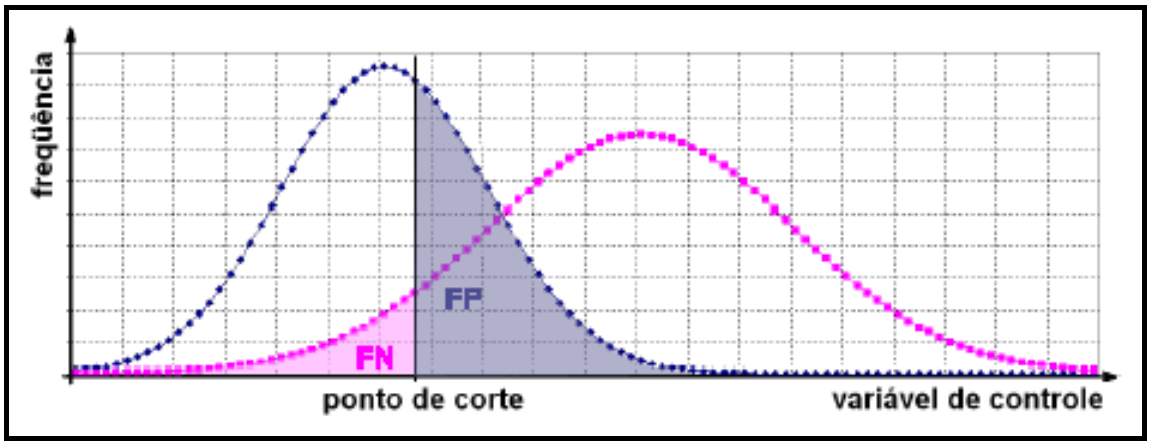


Figura 34: Gráfico com baixo ponto de corte. Fonte: adaptado de (SOVIERZOSKI, ARGOUD e AZEVEDO, 2008).

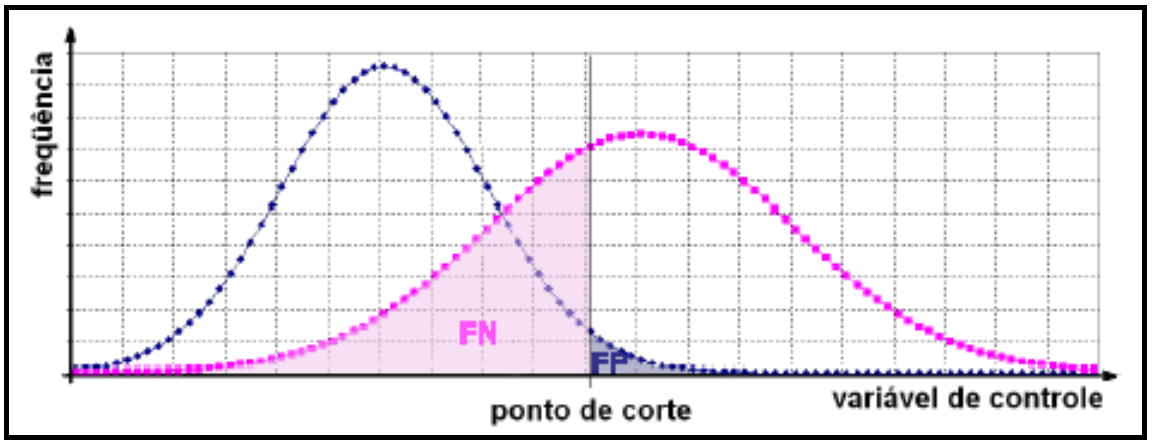


Figura 35: Gráfico com ponto de corte elevado. Fonte: adaptado de (SOVIERZOSKI, ARGOUD e AZEVEDO, 2008).

Conforme (EBERHART e DOBBINS, 1990) a curva ROC apresenta a dependência entre a sensibilidade e a especificidade de um classificador. É um gráfico cartesiano, indicando a fração de *VP* (sensibilidade) no eixo das abscissas e a fração de *FP* (1 – especificidade) no eixo das ordenadas (Figura 36).

Para cada valor da variável de controle no gráfico da Figura 33 corresponde uma combinação de valores de sensibilidade e especificidade. Este par de valores representa um ponto na curva ROC. A linha pontilhada diagonal, na Figura 36, representa um classificador que não consegue discriminar, devido ao percentual de *VP* ser igual ao percentual de *FP*.

Segundo (BROWN e DAVIS, Herbert; 2006) o índice mais importante da análise ROC é o índice *AUC* (do inglês *Area Under the ROC Curve*), assumindo valores entre 0,5 (sem discriminação, sob a linha diagonal tracejada) e 1,0 (discriminação ideal, no canto indicado pela seta azul).

Neste trabalho também foi utilizado o índice *AUC* para a análise do desempenho da metodologia proposta.

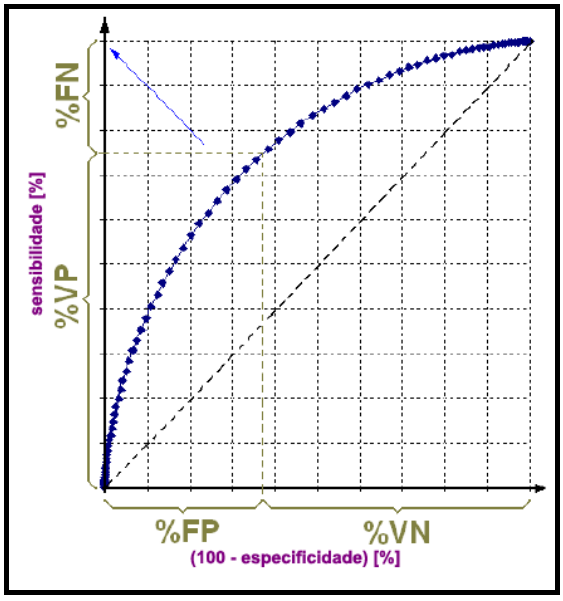


Figura 36: A curva ROC representando a relação entre a sensibilidade e a especificidade do classificador. Fonte: (SOVIERZOSKI, ARGOUD e AZEVEDO, 2008).

3 METODOLOGIA

Este capítulo descreve os procedimentos realizados pela metodologia proposta neste trabalho para detecção de massas em imagens digitais de mamografia. Esta metodologia segue o esquema tradicional de processamento de imagens, apresentado na Seção 2.4, também composta pelas etapas: 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 37 apresenta um diagrama ilustrando essas etapas.

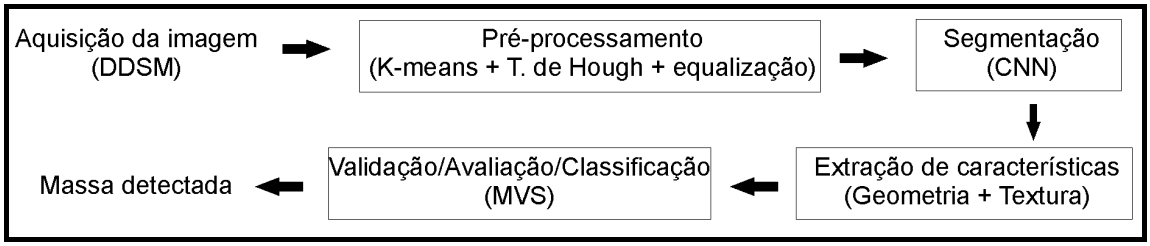


Figura 37: Etapas aplicadas na metodologia.

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 disponível na internet (HEATH *et al*. 1998) (DDSM, 2001).

A etapa de pré-processamento tem o objetivo de facilitar o processamento a ser realizado pelas etapas seguintes através da remoção de elementos indesejáveis usando do algoritmo de agrupamento K-means com duas classes, o filtro de Canny, transformada de Hough e o operador morfológico de erosão matemática e o realce de contraste da imagem através da equalização do histograma da imagem. A etapa de segmentação das

regiões de interesse identifica as áreas da imagem que são suspeitas de conterem massas usando CNN, de forma que as etapas seguintes trabalhem apenas com as regiões relevantes. A etapa de extração de características descreve as regiões de interesse através de suas características de geometria (excentricidade, circularidade, compacidade, desproporção circular e densidade circular) e textura (função K de Ripley, índices de Moran e Geary). Finalmente, a etapa de classificação seleciona as áreas que contém massas e descarta as que contêm não-massas através do classificador MVS.

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

3.1 DDSM

As imagens utilizadas são fornecidas pelo DDSM - *Digital Database for Screening Mammography* (DDSM, 2001) que é um banco de dados público contendo 2.620 casos, fornecidas gratuitamente na internet (HEATH, *et al*. 1998), através dos esforços de alguns 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 (projeções CC e MLO), além de informações extras sobre o exame (data do estudo, idade do paciente, tipo da patologia, quantidade de anomalias, etc.) e sobre a imagem (nome do arquivo, tipo de filme, data de digitalização, tipo do digitalizador, seqüência, *pixels* por linha, bits por *pixel*, marcação, etc.). Todas as informações contidas no DDSM foram fornecidas por especialistas (HEATH, *et al*. 1998).

Neste trabalho foram utilizadas 623 imagens do banco de imagens DDSM. Essas imagens foram selecionadas aleatoriamente, onde a única exigência era que cada imagem deveria possuir apenas uma massa.

3.2 Pré-processamento

Imagens pertencentes à base de imagens DDSM apresentam ruídos e elementos típicos do exame mamográfico que podem interferir no processamento das imagens. 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 internas da mama.

As imagens utilizadas neste trabalho apresentam elementos externos à mama, como o fundo com intensidade próxima de preto, rótulos de marcação e bordas que devem ser removidos para que não influenciem nos resultados das etapas seguintes. Exemplos desses elementos são mostrados na Figura 38.

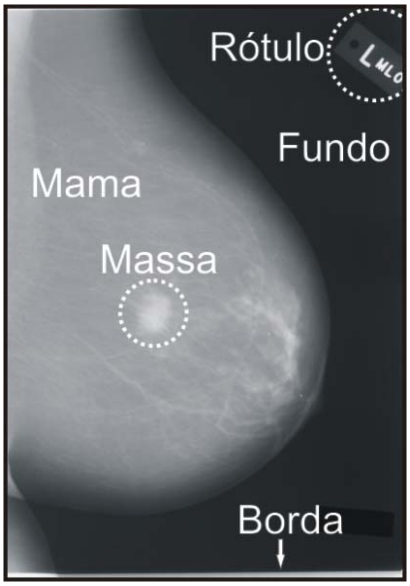


Figura 38: Elementos de uma imagem do DDSM.

A etapa de pré-processamento consiste primeiramente na redução do tamanho da imagem original. Resultando numa imagem com a altura igual a 1024 *pixels*, reduzindo-se proporcionalmente a largura (Figura 39a). Esta

redução se justifica para reduzir o tempo de processamento que se tornaria muito elevado quando se trabalha com a imagem no seu tamanho original.

Após a redução de tamanho, todos os pontos a 30 *pixels* de distância das bordas laterais são removidos. Essa remoção serve para eliminar da imagem, a área entre as bordas do filme da radiografia e os espaços vazios sem o filme.

Em seguida o fundo, com níveis de cinza próximos ao preto, é removido utilizando-se o algoritmo de agrupamento K-means (Seção 2.6.1) para agrupamento dos *pixels* em dois grupos (k=2) de acordo com suas intensidades (Figura 39b). 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 eliminar o grupo contendo os *pixels* de menor intensidade, substituem-se seus valores de intensidade por zero (Figura 39c). A representação visual dos grupos gerados pode ser vista no exemplo da Figura 39.

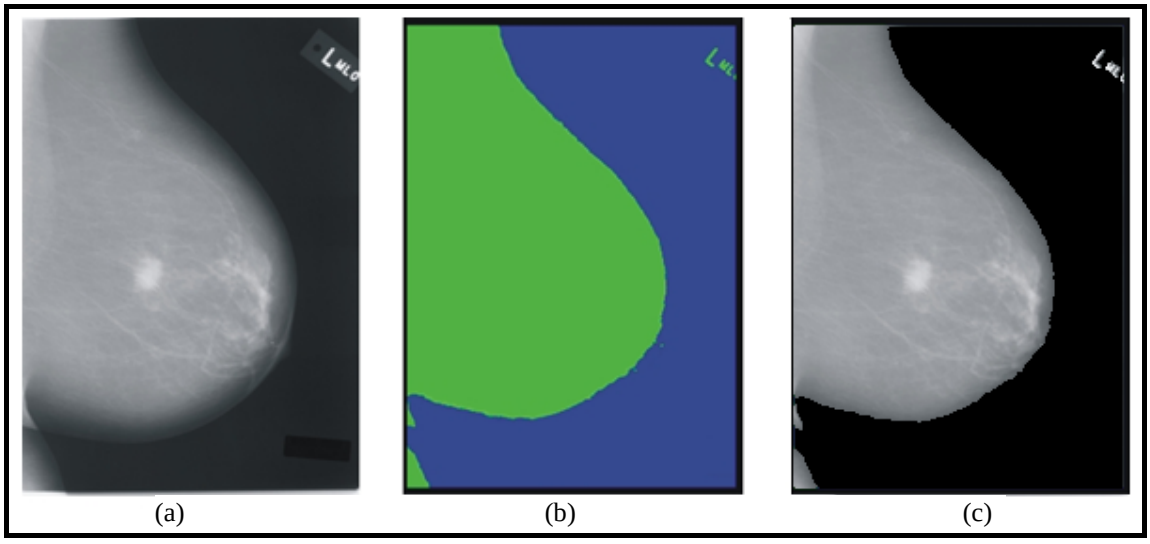


Figura 39: Ilustração da remoção do fundo. (a) Imagem reduzida. (b) Imagem com as bordas removidas e realizado um agrupamento com o *K-means* (*k*=2). (c) Imagem com o fundo removido.

Na imagem resultante ainda existem objetos externos à mama como rótulos de identificação do exame e algum ruído. Esses objetos são removidos através do algoritmo de crescimento de região (Seção 2.4.2) que separa na imagem as regiões não conexas (Figura 40). Seleciona-se apenas a região de maior área, ou seja, a mama.

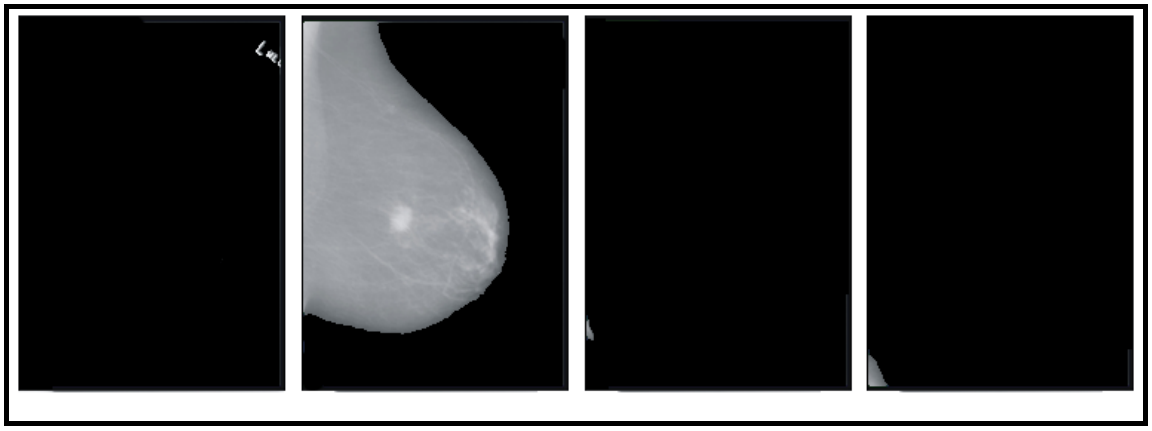


Figura 40: Objetos não conexos separados pelo algoritmo de crescimento de região.

Para finalizar a etapa de pré-processamento, realiza-se um realce de contraste da imagem através da equalização do histograma da imagem (Seção 2.4.1), com isso melhora-se a visualização das estruturas internas da mama (Figura 41). Foram testadas outras formas de realce, como equalização do histograma usando a imagem total e realce através de ANCE (*Adaptive Neighborhood Contrast Enhancement*) (STRICKLAND, 2002). Porém a técnica de realce por equalização de histograma apenas na área que contém a mama, não se considerando o fundo preto, foi a técnica que possibilitou uma melhor segmentação.

As imagens MLO geralmente contêm o músculo peitoral, que está associado a altos níveis de cinza, o que interfere na distribuição das intensidades dos *pixels* após a equalização do histograma da imagem. Um exemplo é apresentado na Figura 42.

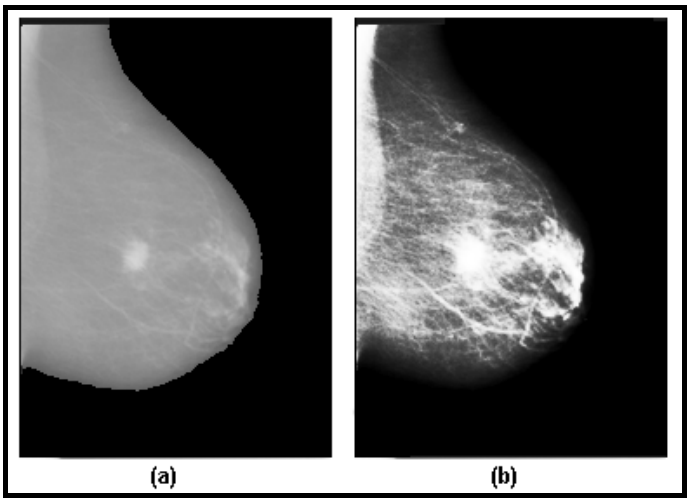


Figura 41: Realce de contraste por equalização do histograma. (a) Imagem sem realce. (b) Imagem com realce.

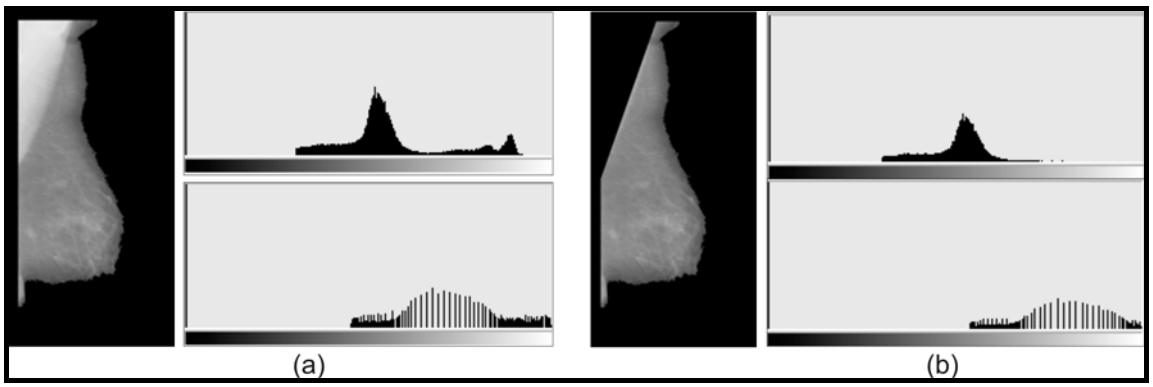


Figura 42: Comparação de histogramas: (a) Imagem com o músculo peitoral. Ao lado, os histogramas original (cima) e equalizado (baixo). (b) Imagem sem o músculo peitoral. Ao lado os histogramas original (cima) e equalizado (baixo).

Assim, todas as imagens com incidência MLO foram submetidas a um novo pré-processamento para remoção do músculo peitoral. Inicialmente localiza-se o lado em que se encontra o músculo peitoral. Esta localização é feita dividindo-se a imagem na metade de sua largura, então se calcula a média da intensidade dos *pixels* de cada metade. A metade que contiver a maior média será a que contém a mama. Todos os *pixels* abaixo da metade da altura da imagem e na metade oposta a este lado são removidos (Figura 43). Em seguida, aplica-se o filtro de Canny (Seção 2.4.3) para detecção de bordas. Remove-se então, através do operador morfológico de erosão (Seção 2.4.4),

todas as bordas detectadas que não estão na direção da borda associada ao músculo peitoral (Figura 44). Todos os segmentos menores que 5 *pixels* e maiores que 15 são removidos (Figura 43c), reduzindo a quantidade de *pixels* e o número de segmentos de reta a serem analisados. Em seguida, usando a transformada de Hough (Seção 2.4.5) encontra-se a reta que melhor representa a borda do músculo peitoral (Figura 43d). Todos os *pixels* da borda do músculo até a reta são removidos.

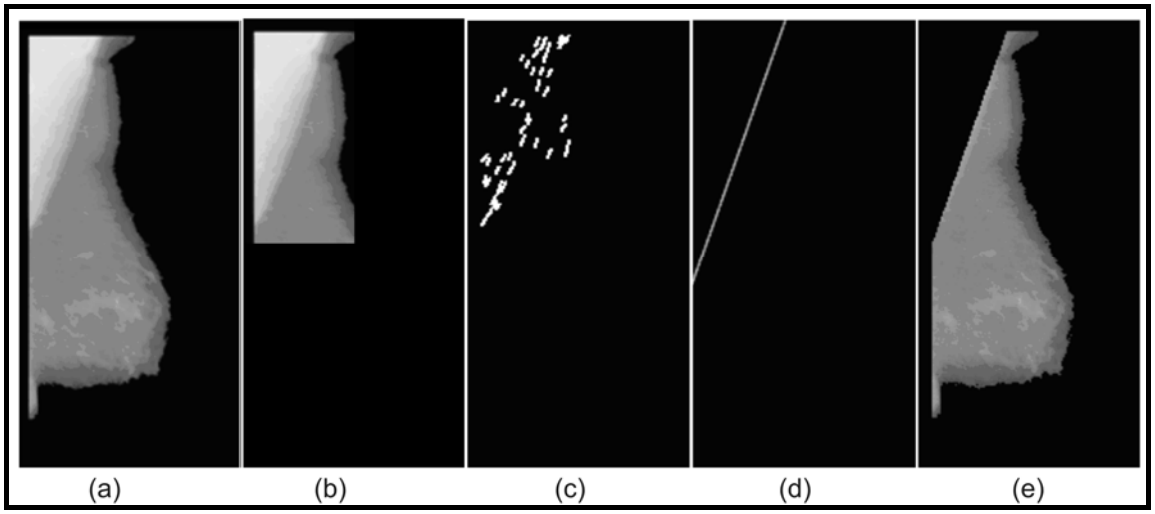


Figura 43: Remoção do músculo peitoral. a) Imagem MLO. b) Área do músculo localizada. (c) Bordas de mesmo sentido que o músculo peitoral localizadas pelo filtro de Canny e filtradas por erosão matemática. d) Reta detectada pela transformada de Hough; e) Imagem sem o músculo peitoral.

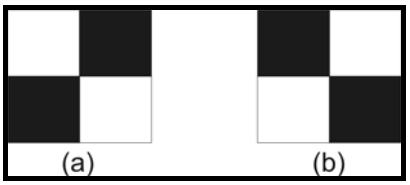


Figura 44: Elementos estruturantes usados para erosão matemática. (a) Elemento estruturante utilizado para músculo peitoral na direita. (b) Elemento estruturante utilizado para músculo peitoral na esquerda.

3.3 Segmentação

A etapa de segmentação tem o objetivo de identificar as regiões da mama com maiores possibilidades de conterem massas. Neste trabalho a segmentação é feita através de Redes Neurais Celulares (Seção 2.6.2.3), utilizando dois *templates* para gerar os candidatos a massa.

O primeiro é o *template Textudil* (CHUA e ROSKA, 2004), que consegue separar massas, mas tem o efeito colateral de incluir *pixels* não pertencentes à massas.

O segundo *template*, *Blur* (ZARÁNDY, *et al*. 1994), evita a inclusão de *pixels* extras nos candidatos, porém pode remover muitos *pixels* da massa. As configurações dos *templates* são mostrados nas Figuras 45 e 46. Onde *A* é o operador de *feedback*, *B* é o operador de entrada sináptica e *Z* é o limiar.

Os limiares dos *templates Textudil* e *Blur* foram determinados empiricamente, sendo utilizados os valores que apresentaram melhor desempenho, ou seja, $Z = 4.5$ e $Z = 1.0$ respectivamente.

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \qquad Z = 4.5$$

Figura 45: Configuração do *template Textudil*.

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}, \qquad Z = 1.0$$

Figura 46: Configuração do *template Blur*.

A Figura 47 contém um exemplo de segmentação utilizando os *templates Textudil* e *Blur* respectivamente.

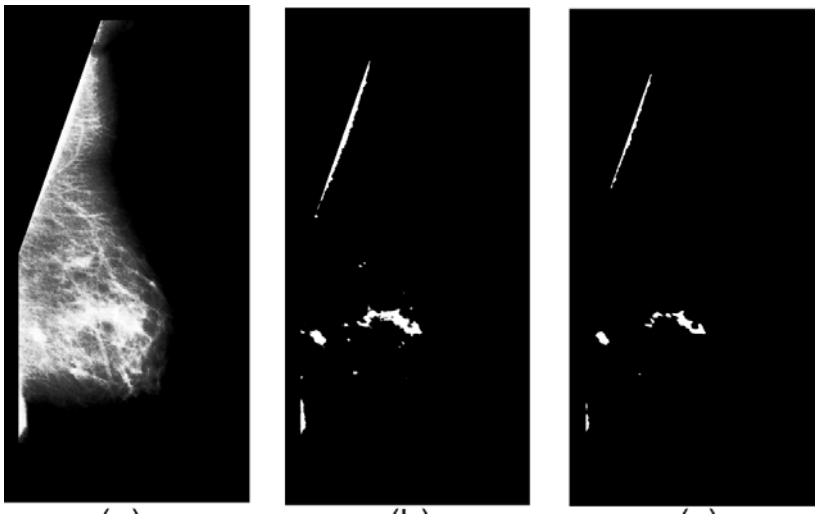


Figura 47: Segmentação em uma imagem pré-processada (a) através de CNN usando os *templates Textudil* (b) e *Blur* (c).

Após a segmentação, todos os candidatos menores que 15x15 *pixels* ou maiores que 300x300 *pixels* são excluídos. Assim evita-se que candidatos muito pequenos ou muito grandes passem para as etapas posteriores. Esses valores foram baseados em trabalhos anteriores, conforme (NUNES, 2009) e (MARTINS, *et al*. 2009).

A seguir a imagem resultante da segmentação usando os *templates Textudil* e *Blur* são somadas. Essa soma serve para incluir candidatos perdidos por um dos *templates* e para evitar que candidatos sejam repetidos. Após essa soma, separam-se todos os candidatos, por crescimento de região (Seção 2.4.2), e eliminam-se o que não obedecem aos critérios de tamanho (Figura 48).

Deve-se observar que os objetos encontrados não possuem textura. No final desta etapa, a textura é incluída no objeto, copiando-se os *pixels* da área correspondente na imagem pré-processada (melhorada), para análise posterior.

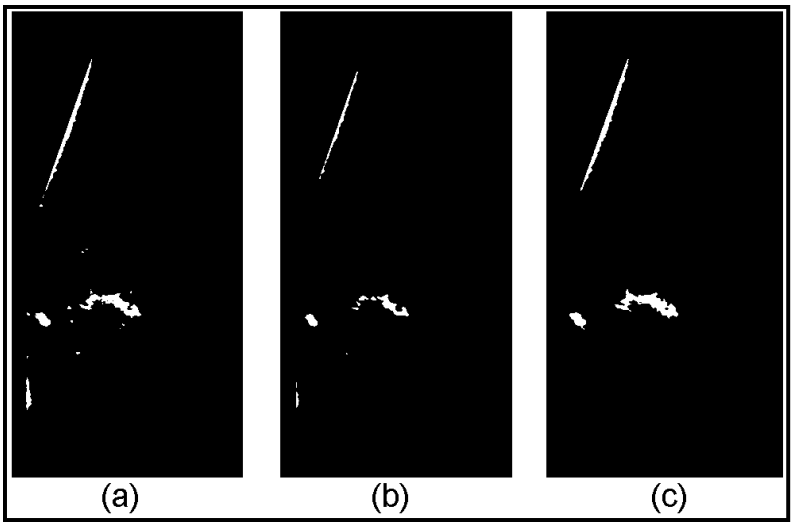


Figura 48: Objetos somados e filtrados. (a) Objetos selecionados pelo *template Textudil*. (b) Objetos selecionados pelo *template Blur*. (c) Objetos somados e filtrados.

3.4 Extração de características

Nesta etapa, o objetivo é extrair medidas descritivas das regiões de interesse segmentadas na etapa anterior para formar os vetores de características que serão utilizados 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 (Seção 2.4.7): excentricidade, circularidade, compacidade, desproporção circular e densidade circular. O procedimento de extração dessas medidas não leva em conta as intensidades dos *pixels* das regiões de interesse.

A textura das regiões de interesse é descrita através (Seção 2.4.8) da Função K de Ripley (na forma local), Índice de Moran e Geary.

Para a análise utilizando a função K de Ripley, calcula-se inicialmente o centro de massa de cada candidato. Em seguida, encontra-se o maior raio (R) possível que se inicia no centro de massa do candidato e vai até o ponto mais distante da borda do mesmo. A análise é feita em duas regiões. A primeira corresponde ao círculo (área interna) com o mesmo centro de massa

do candidato e que possui um raio *r=R/2*. A segunda região corresponde ao anel que possui a borda externa a uma distância *R* do centro de massa do candidato e com a borda interna a uma distância *r* do mesmo centro (Figura 49).

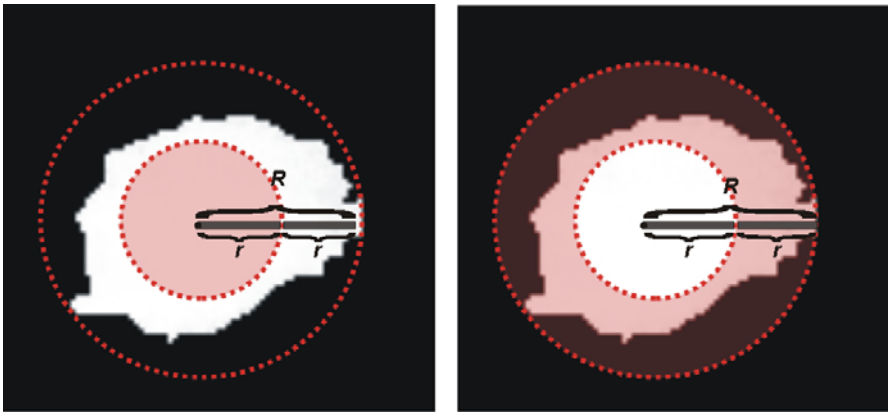


Figura 49: Exemplo da utilização da função K de Ripley em um candidato a massa.

Para o cálculo dos Índices de Moran e Geary foram utilizados os *pixels* vizinhos de 1 até 10 unidades de distância de um *pixel* analisado, e em quatro direções, correspondendo aos ângulos 0º, 45º, 90º, 135º.

A Figura 50 contém ilustrações da análise de um candidato a massa através dos Índices de Moran e Geary.

Essa abordagem de utilizar quatro direções e os vizinhos que variam de 1 até 10 unidades de distância do *pixel* analisado se justifica pelo fato da análise tradicional dos Índices de Moran e Geary ser computacionalmente inviável para imagens onde os candidatos apresentam maior área. Tal modificação permite uma análise localizada por *pixel* e reduz significativamente o tempo de processamento necessário.

É importante ressaltar que para cada tipo de descritor de textura, foram utilizados vários níveis de quantização da imagem. Foram utilizados 256, 128, 64, 32, 16 e 8 níveis de cinza para cada candidato a massa. Essas quantizações visam aumentar os relacionamentos de textura possíveis.

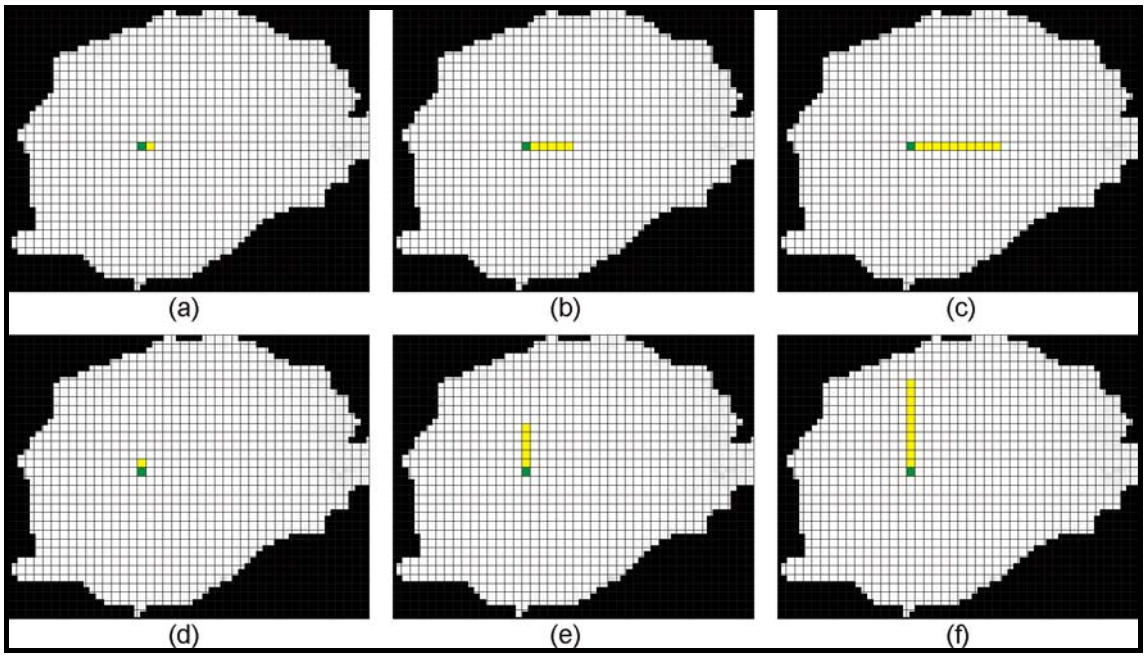


Figura 50: Ilustrações da análise de textura através dos Índices de Moran e Geary. O *pixel* verde representa um *pixel* sendo analisado e os *pixels* de amarelo correspondem a sua vizinhança com distâncias 1 (a, d), 5 (b, e) e 10 (c, d); e direções 0º (a, b, c) e 90º graus (d, e, f).

Ao final do processo de extração de características cada região de interesse é representada por um vetor contendo informações sobre a geometria e textura. Mais precisamente temos a geometria contribuindo com 5 características; Ripley com [(256+128+64+32+16+8) quantizações] X 2 raios = 1008 características; Moran com [(256+128+64+32+16+8) quantizações] X 4 direções X 10 vizinhanças = 240 características; Geary com [(256+128+64+32+16+8) quantizações] X 4 direções X 10 vizinhanças = 240 características.

Foram feitas várias combinações de características e analisados seus desempenhos através do classificador MVS utilizando-se todas as características e também as selecionadas pelo método de seleção de variáveis *stepwise*.

3.5 Seleção de características

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 2.5, 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.

3.6 Classificação

Na etapa de classificação, o conjunto de dados, é dividido em dois conjuntos, de treinamento e teste. Este trabalho utilizou a técnica de validação cruzada *leave-N-out* (Seção 2.7) por ser independente da proporção da divisão entre treino e teste.

Primeiramente, a amostra foi separada em dois grupos: massas e não-massas. Em seguida, cada grupo foi dividido aleatoriamente em 10 subconjuntos. Onde um subconjunto é escolhido para teste e os subconjuntos restantes são utilizados para treinamento. Repetiu-se este processo até que todos os subconjuntos tivessem sidos testados.

Neste trabalho, o classificador MVS foi utilizado com núcleo radial e parâmetros padrão (C=1 e y=0,5). Devido a proporção de não-massas selecionadas na etapa de segmentação ser aproximadamente 6 vezes maior que a quantidade de massas, atribuiu-se um peso maior ao treinamento das massas. Isso significa que no treinamento, a penalidade por errar uma massa é maior que errar uma não-massa. Inicialmente testou-se o peso 6, porém observou-se que havia um grande desequilíbrio entre sensibilidade e especificidade. Um bom equilíbrio entre esses dois índices foi conseguido utilizando-se os pesos w_{massa}= 9 e $w_{não\text{-}massa}$=1.

Foi feita a análise do desempenho da classificação utilizando-se as características totais, ou seja, sem redução e seleção de características, e com as características selecionadas pelo método de seleção *stepwise*.

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

4 RESULTADOS

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

Inicialmente foram selecionadas, de forma aleatória, 700 imagens, mas devido à restrição de que cada imagem deveria conter apenas uma massa, 77 imagens foram excluídas da análise da metodologia proposta, sobrando uma amostra de 623 imagens extraídas do DDSM.

A partir dessas imagens, a etapa de segmentação das regiões de interesse selecionou um total de 3871 regiões suspeitas de conterem anormalidades, sendo 566 massas de fato e de 3305 não-massas.

Em 57 imagens das 623 iniciais o procedimento de segmentação falhou em incluir as massas presentes no conjunto de regiões de interesse, o que representa 9,15% dos casos. Este fato sugere, que os parâmetros utilizados nos *templates* devem ser otimizados para se obter um desempenho melhor da segmentação.

As 566 massas segmentadas corretamente representam 90,85% dos casos. Esses resultados demonstram que a etapa de segmentação apresenta uma boa sensibilidade, mas produz muitos falsos positivos, mais precisamente 5,30 por imagem (3305 regiões que não contém massas, segmentadas como regiões de interesse nas 623 imagens), os quais se espera que sejam eliminados durante a etapa de classificação.

Testes foram realizados com e sem redução de características com o objetivo de se fazer uma análise do comportamento da classificação. Nesta etapa, os candidatos segmentados foram divididos em 10 grupos. Essa divisão tem o objetivo de se utilizar a técnica *leave-N-out*, onde todos os grupos são treinados e testados, evitando-se assim testes tendenciosos. A seleção dos objetos em cada grupo foi feita aleatoriamente a partir do total de regiões segmentadas. As próximas subseções descrevem cada uma dessas abordagens e discute os resultados obtidos.

4.1 Testes sem redução de variáveis

Na primeira abordagem de testes, utilizaram-se as características de geometria e textura para descrever as regiões a serem classificadas. Mais precisamente temos a geometria contribuindo com 5 características; Ripley com [(256+128+64+32+16+8) quantizações] X 2 raios = 1008 características; Moran com 6 quantizações X 4 direções X 10 vizinhanças = 240 características; Geary com 6 quantizações X 4 direções X 10 vizinhanças = 240 características.

A Tabela 1 mostra a média dos indicadores de desempenho obtidos através do método *leave-N-out* com 10 grupos. Inicialmente cada descritor foi utilizado de forma isolada, com o objetivo analisar o comportamento desse descritor sem receber influência dos demais. Como os índices de Moran e Geary são muito parecidos (medem a correlação espacial), um novo teste foi realizado com a combinação de ambos.

Tabela 1**:** Resultados dos testes sem redução de características.

Métodos	Sensib. (%)	Especif. (%)	Acurácia (%)	FP/i	FN/i	Overlay	AUC
Geometria	78,55	61,45	63,95	2,25	0,21	0,320	0,753
Ripley	87,25	56,01	60,58	2,57	0,13	0,306	0,738
Moran	75,65	55,32	58,29	2,61	0,24	0,381	0,698
Geary	69,36	63,00	63,93	2,16	0,31	0,392	0,704
Moran+Geary	78,57	70,17	71,43	1,74	0,21	0,362	0,792

Pode-se observar na Tabela 1, que a análise através da função K de Ripley possui maior sensibilidade (87,25%), porém apresenta uma baixa especificidade (56,01%). Os índices de Moran e Geary quando usados de forma separada apresentam um baixo desempenho, no entanto, quando combinados apresentaram o melhor desempenho geral. É importante observar que o índice Overlay apresenta abaixo desempenho, mas deve-se salientar que o especialista ao marcar a área contendo a massa na mamografia, marca uma área geralmente maior que a massa propriamente dita.

De forma geral, a abordagem sem redução de características apresentou um baixo desempenho na classificação dos candidatos a massa.

Isto ocorre porque algumas características são irrelevantes, redundantes ou apresentarem níveis elevados de ruído.

Outras desvantagens dessa abordagem estão no tempo necessário para o processamento da classificação e na grande quantidade de memória exigida, o que impede maiores combinações.

A Figura 51 contém as curvas ROC geradas.

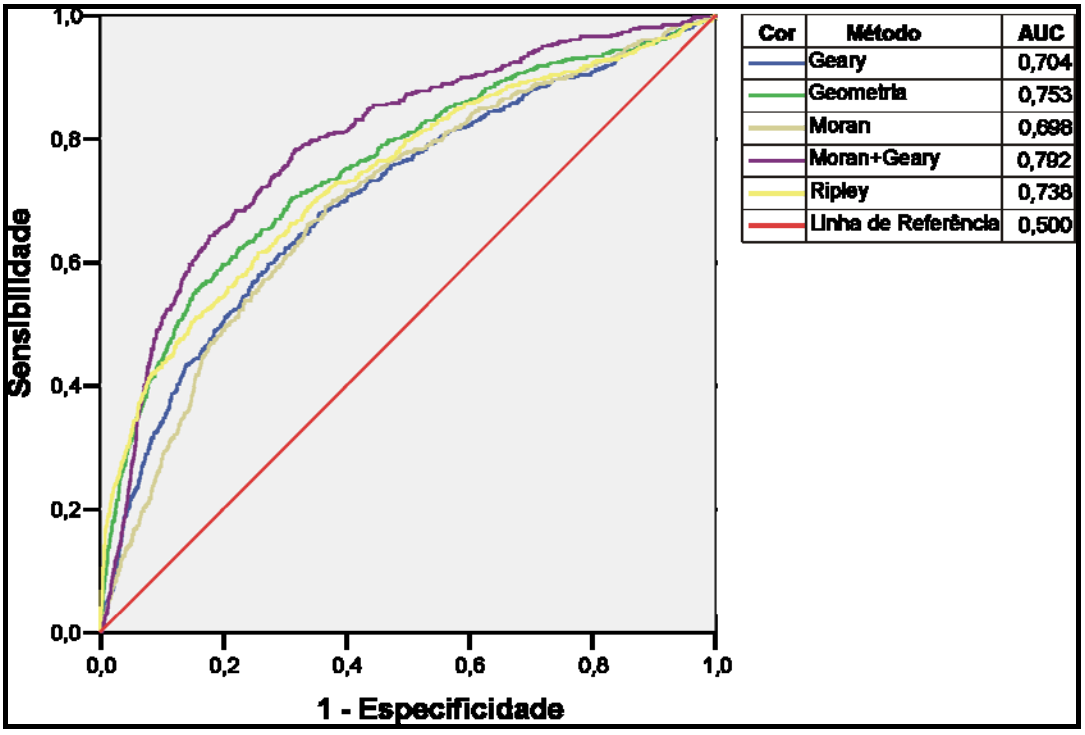


Figura 51: Curvas ROC geradas com a análise usando todas as características.

Observa-se na Figura 51 que a combinação entre os índices de Moran e Geary apresenta a maior área (AUC = 0,792) e o índice de Moran quando utilizado de forma isolada apresenta a menor (AUC = 0,698).

4.2 Testes com redução de variáveis

Na segunda abordagem de testes, os vetores de características foram formados utilizando-se apenas as características selecionadas pelo método de

redução de variáveis *stepwise*, obtendo-se as seguintes quantidades: Geometria – 4 características (Excentricidade, circularidade, compacidade, densidade circular); Ripley – 23 características (Tabela 2); Moran – 12 características (Tabela 3); Geary – 7 características (Tabela 4).

Tabela 2: Variáveis selecionadas da função K de Ripley.

Quantização	Anel	Níveis
8	1	4, 7
16	0	0
	1	14
32	0	31
64	1	56, 58
128	0	0, 120, 127
	1	116, 127
256	0	193, 219, 235, 241, 242, 243, 253
	1	248, 249, 251, 252

Tabela 3: Variáveis selecionadas do índice de Moran.

Quantização	Direção	Distância
8	90º	2, 10
32	90º	1
	135º	5
64	135º	4
128	90º	2, 7
	135º	6
256	0º	1, 6
	135º	8, 9

Tabela 4: Variáveis selecionadas do índice de Geary.

Quantização	Direção	Distância
32	90º	1, 4
	135º	10
64	0º	2, 4
128	0º	10
256	135º	1

A Tabela 5 mostra a média dos indicadores de desempenho obtidos através do método *leave-N-out* com 10 grupos. Cada descritor de característica foi analisado separadamente, e também através da combinação dos mesmos.

Tabela 5: Resultados dos testes com redução de características.

Métodos	Sensib. (%)	Especif. (%)	Acurácia (%)	FP/i	FN/i	Overlay	AUC
Geometria	77,64	62,15	64,42	2,21	0,22	0,321	0,753
Ripley	81,23	68,50	70,37	1,84	0,19	0.306	0,803
Moran	**77,42**	**51,51**	**55,3**	**2,83**	**0,23**	**0,381**	**0,696**
Geary	70,66	61,99	63,25	2,22	0,29	0,390	0,708
C(1,2)	81,24	74,56	75,54	1,49	0,19	0,303	0,819
C(1,3)	83,39	67,42	69,75	1,90	0,17	0,337	0,792
C(1,4)	76,85	75,80	75,96	1,41	0,23	0,349	0,796
C(1,2,3)	79,64	79,61	79,62	1,19	0,20	0,324	0,850
C(1,2,4)	76,98	82,73	81,90	1,01	0,23	0,325	0,846
C(1,3,4)	85,41	75,61	77,07	1,42	0,15	0,329	0,846
C(1,2,3,4)	**80,00**	**85,68**	**84,62**	**0,84**	**0,20**	**0,325**	**0,870**
C(2,3,4)	81,75	81,86	81,86	1,06	0,18	0.327	0,861
C(3,4)	84,41	62,68	65,89	2,18	0,16	0,353	0,794

Legenda: C-combinação, (1)-Geometria, (2)-Ripley, (3)-Moran, (4)-Geary

Observa-se na Tabela 5, que a combinação da geometria e Índices de Moran e Geary possui a maior sensibilidade (85,41%), porém apresenta uma baixa especificidade (75,61%). O melhor desempenho geral é encontrado através da combinação de todas as características (linha verde), no que se reflete na maior acurácia (84,62%) e na maior área sob a curva ROC (0,870). Outra observação é através da análise utilizando apenas textura, pois foi a que apresentou um bom desempenho e maior equilíbrio entre sensibilidade (81,75%), especificidade (81,86%) e acurácia (81,86%), e a segunda maior área sobre a curva ROC (0,861). A análise de pior desempenho (linha vermelha) foi obtida através do índice de Moran, pois apresenta a menor área sob a curva ROC (0,696) e a menor acurácia (55,3%).

Comparando-se os itens da Tabela 1 com seus correspondentes na Tabela 5, percebe-se um ligeiro aumento no desempenho médio do classificador (Tabelas 6 e 7). Este fato se deve a remoção de ruídos, redundâncias e irrelevâncias nos dados, proporcionadas pelo algoritmo de seleção de características *stepwise*.

Tabela 6: Desempenho médio do classificador sem seleção de características.

Métodos	Sensib. (%)	Especif. (%)	Acurácia (%)	FP/i	FN/i	Overlay	AUC
Geometria	78,55	61,45	63,95	2,25	0,21	0,32	0,7
Ripley	87,25	56,01	60,58	2,57	0,13	0,306	0,665
Moran	75,65	55,32	58,29	2,61	0,24	0,381	0,654
Geary	69,36	63	63,93	2,16	0,31	0,392	0,665
Moran+Geary	78,57	70,17	71,43	1,74	0,21	0,362	0,747
Média	**77,876**	**61,19**	**63,636**	**2,266**	**0,22**	**0,3522**	**0,6862**
Desvio Padrão	**6,447**	**6,026**	**4,971**	**0,353**	**0,064**	**0,037**	**0,038**

Tabela 7: Desempenho médio do classificador com as características selecionadas pelo algoritmo *stepwise*.

Métodos	Sensib. (%)	Especif. (%)	Acurácia (%)	FP/i	FN/i	Overlay	AUC
Geometria	77,64	62,15	64,42	2,21	0,22	0,321	0,699
Ripley	81,23	68,5	70,37	1,84	0,19	0.306	0,749
Moran	77,42	51,51	55,3	2,83	0,23	0,381	0,644
Geary	70,66	61,99	63,25	2,22	0,29	0,39	0,662
Moran+Geary	84,41	62,68	65,89	2,18	0,16	0,353	0,644
Média	**78,272**	**61,366**	**63,846**	**2,256**	**0,218**	**0,36125**	**0,6796**
Desvio Padrão	**5,135**	**6,139**	**5,488**	**0,358**	**0,049**	**0,031**	**0,045**

A abordagem com redução de características possibilitou a análise da classificação dos candidatos através de combinação das características extraídas dos candidatos, com a vantagem de ser uma análise rápida e requerer menor quantidade de memória para seu processamento.

Como se pode perceber nas Tabelas 1 e 5, os índices *overlay* ficaram abaixo de 0,40, indicando que, de forma geral, os objetos classificados como massa possuem uma área menor que a área informada no DDSM. Fato explicado pelo maior área informada pelo especialista do que a área real da massa.

A Figura 52 contém as curvas ROC geradas. Nela observa-se que a combinação com todos os descritores apresenta a maior área (AUC = 0,827), e o índice de Moran quando utilizado de forma isolada ou combinado com o índice de Geary apresenta a menor área (AUC = 0,644).

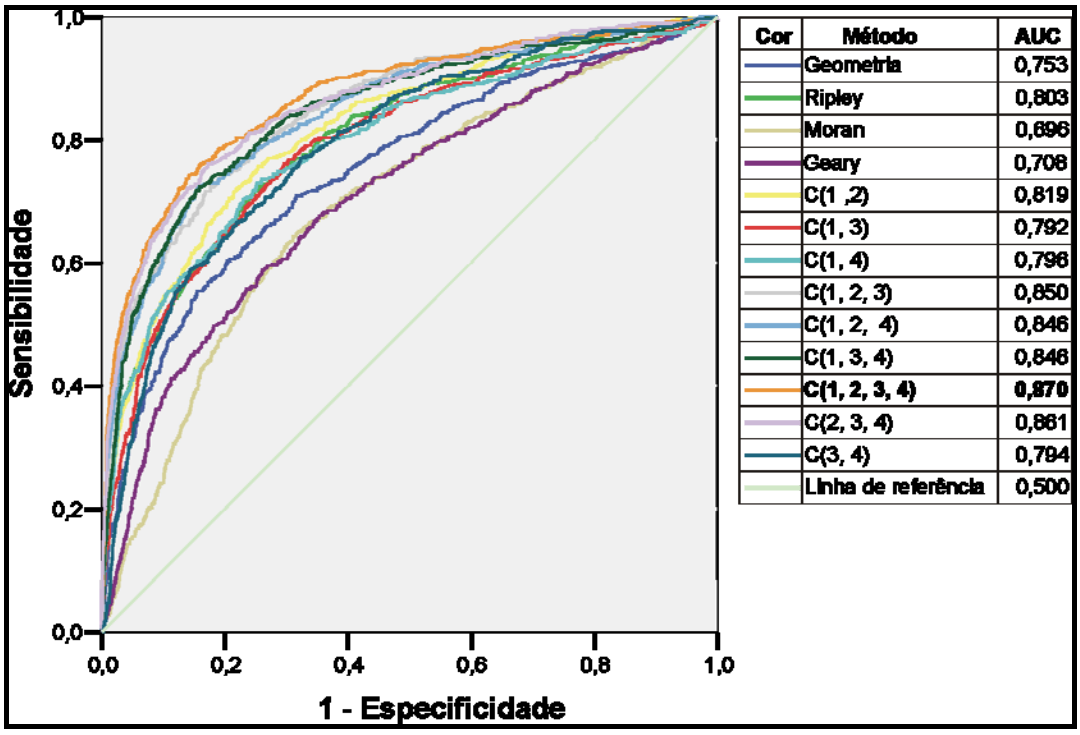


Figura 52: Curvas ROC geradas com a análise usando o algoritmo de *stepwise*. Na legenda: C-combinação, (1)-Geometria, (2)-Ripley, (3)-Moran, (4)-Geary.

4.3 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.

4.3.1 Primeiro Caso: Detecção Correta

O primeiro caso, apresentado na Figura 53, mostra a seqüência de passos realizados para a detecção correta de uma massa em uma imagem de mamografia. A Figura 53a mostra a imagem original da mamografia em questão, tal como se apresenta na base do DDSM.

No pré-processamento, produz-se a imagem da Figura 53b, na qual se pode observar que os objetos externos à mama, foram removidos. Também é possível observar os efeitos da equalização do histograma realizado para realçar as estruturas internas da mama.

A próxima etapa (Figura 53c) realiza a segmentação das regiões de interesse através de CNN usando os *templates Blur* e *Textudil*.

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 53d apresenta, em azul, a região classificada como massa.

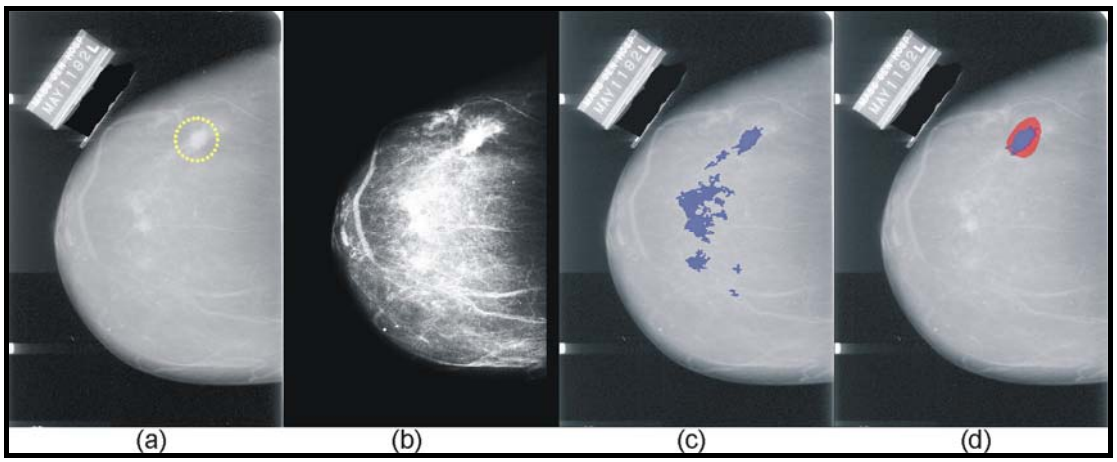


Figura 53: Imagens do estudo de caso 1. (a) Imagem original. Em amarelo está a área contendo a massa. (b) Imagem pré-processada. (c) Em azul estão as regiões de interesse selecionadas na segmentação. (d) Em azul está a massa localizada pela metodologia e de vermelho a área informada no DDSM.

Para verificar o êxito da detecção, a marcação da localização correta da massa obtida a partir das informações contidas no DDSM, é impressa em vermelho sobre a imagem resultante.

4.3.2 Segundo Caso: Falha na classificação

O segundo caso, apresentado na Figura 54, mostra a mesma sequência de passos descritos na seção anterior: imagem original (Figura 54a); resultado do pré-processamento (Figura 54b); candidatos selecionados em azul (Figura 54c) e o resultado da classificação, mostrando em azul a região classificada pela MVS como massa (Figura 54d).

Como a marcação em vermelho indica a localização correta da massa, obtida a partir das informações disponíveis no DDSM, pode-se observar nas Figura 54c e 54d 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 uma não-massa.

Esse caso ilustra a ocorrência de um falso negativo (massa classificada como não-massa) e um falso positivo (não-massa classificada como massa).

Entre os possíveis fatores relacionados a essa falha está o fato de que o candidato selecionado como massa possuir uma forma mais arredondada.

Já o candidato selecionado que realmente corresponde à massa possui uma forma mais curvilínea e alongada, o que causou sua reprovação. Neste caso, a geometria foi mais relevante que a textura.

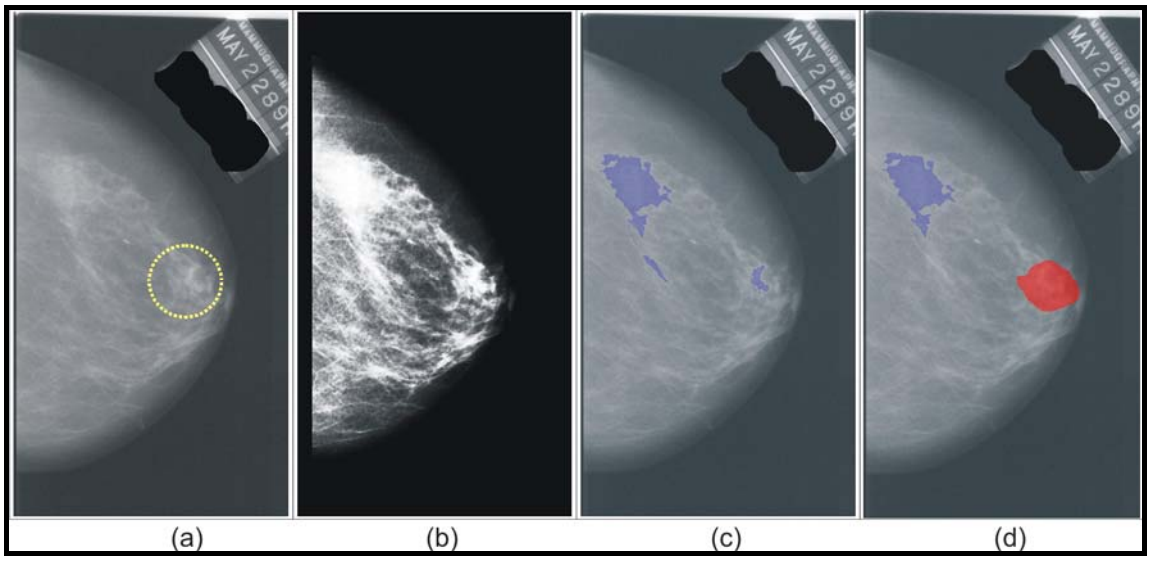


Figura 54: Imagens do estudo de caso 2. (a) Imagem original. Em amarelo está a área contendo a massa. (b) Imagem pré-processada. (c) Em azul as regiões de interesse selecionadas na segmentação. (d) Em azul está o candidato erroneamente selecionado como massa e em vermelho está região contendo massa informada no DDSM.

4.3.3 Terceiro Caso: Falha na segmentação

O terceiro caso apresenta um exemplo de falha da metodologia em realizar uma segmentação adequada.

A Figura 55 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 55c), como se pode observar pela marcação da localização correta da massa, em vermelho, na Figura 55d. 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 pode está relacionada com o critério de tamanho imposto para se considerar uma região segmentada como uma região de interesse. Ou seja, a região pode ter sido segmentada,

mas ficou unida com uma estrutura extra muito grande, ou a segmentação removeu *pixels* em excesso do candidato, deixando-o muito pequeno.

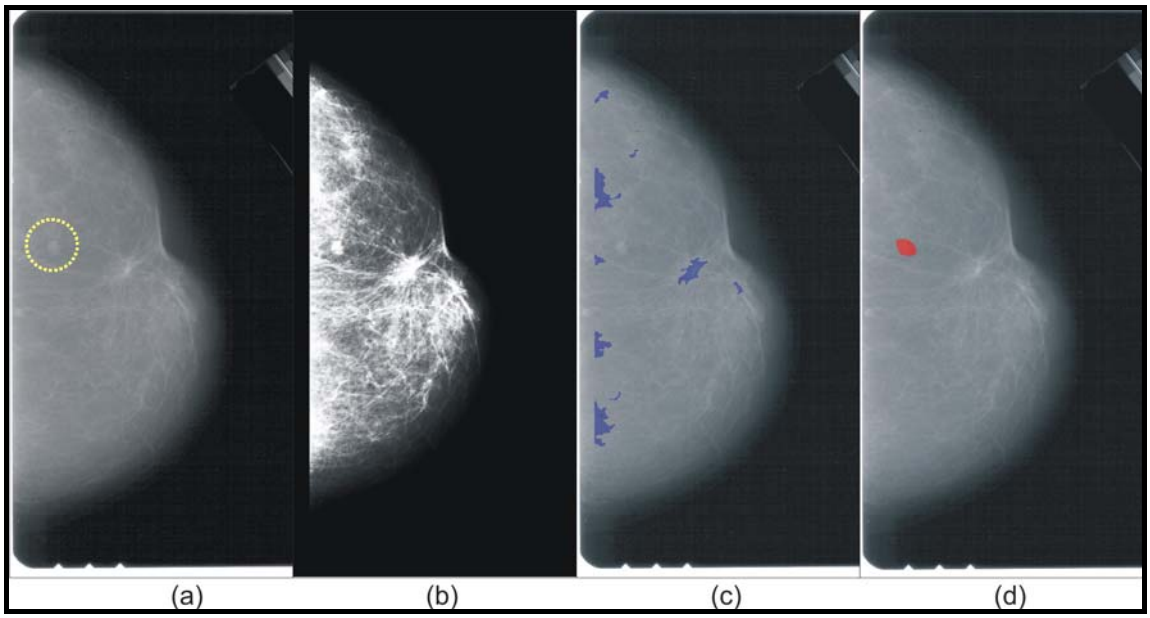


Figura 55: Imagens do estudo de caso 3. (a) Imagem original. Em amarelo está a área contendo a massa. (b) Imagem pré-processada. (c) Regiões de interesse selecionadas na segmentação. (d) Em vermelho está a região contendo massa informada pelo DDSM.

4.4 Comparação com outros trabalhos

Esta seção tem o objetivo de fazer comparações entre a metodologia proposta neste trabalho e as propostas em outros que envolvem detecção de massas em imagens de mamografias, e se ter uma noção da qualidade do trabalho aqui apresentado.

É importante ressaltar que, para uma comparação justa das metodologias citadas, seria necessária a utilização de uma mesma base de imagens, as mesmas imagens deveriam ser comuns a todos os trabalhos. Além disso, seria necessário padronizar alguns parâmetros, como resolução, bits por *pixel*, etc. Outro fator que deveria ser comum aos trabalhos é a amostra

utilizada, pois todas as metodologias deveriam conter os mesmos dados utilizados na fase de treinamento e de teste.

A Tabela 8 contém uma comparação dos resultados de trabalhos envolvendo detecção de massas em imagens mamográficas.

Tabela 8: Comparação entre metodologias de detecção de massas.

Metodologia	Base	Sensib. (%)	Espec. (%)	Acurácia (%)	FP/i	FN/i	AUC	Overlay
NUNES, 2009	DDSM	83,24	84,14	83,94	0,55	0,17	---	---
MARTINS, *et al*. 2009	DDSM	---	---	89,30	0,93	0,02	---	---
NASCIMENTO e RAMOS, 2008	DDSM	94,40	96,90	91,80	---	---	---	---
HASSANIEN, 2007	MIAS	---	---	98,46	---	---	---	---
KOM, TIEDEU e KOM, 2007	Proprietária	95,91	---	---	---	---	0,946	---
ELTONSY, TOURASSI e ELMAGHRABY, 2007	DDSM	81,00	---	---	0,6	---	---	---
BELLOTTI. *et al*. 2006	MAGIC-5	82,00	---	---	2,8	---	0,862	---
TÓTH, TAKÁCS e PATAKI, 2005	DDSM	95,10	---	---	4,3	---	---	---
SERHAT, ONUR e YILMAZ, 2005	MIAS	81,00	---	---	0,33	---	---	---
YING, XINBO e JIE, 2007	DDSM	90,60	---	---	3,60	---	---	---
IREANEUS, THAMARAI, 2008	MIAS	---	---	92,00	---	---	---	---
TIMP, VARELA e KARSSEMEIJER, 2007	Proprietária	---	---	---	---	---	0,770	---
METODOLOGIA PROPOSTA	**DDSM**	**80,00**	**85,68**	**84,62**	**0,84**	**0,20**	**0,870**	**0,325**

Observa-se na Tabela 8 que a base de dados DDSM é amplamente utilizada entre as metodologias de detecção de massas.

Na tabela 8, os valores informados sobre a metodologia proposta, são os valores obtidos através da análise envolvendo todos os descritores (geometria e função k de Ripley e os índices de Moran e Geary), pois entre todos os testes, esta análise que apresentou a maior acurácia (84,62%), a menor taxa de falsos positivos por imagem (0,84), a maior área sob a curva ROC (0,870) e a maior especificidade (85,68%).

A grande maioria dos trabalhos citados utiliza a sensibilidade como índice de desempenho de detecção. A metodologia proposta atingiu uma boa sensibilidade de 80,00%. Este índice influencia inversamente o índice de falsos negativos por imagem (FN/i), onde na metodologia proposta foi igual a 0,2.

A grande maioria dos trabalhos não utiliza a especificidade como métrica de desempenho. Isso evita sabermos qual a taxa de tecidos sadios foram classificados como massa. Pelas informações obtidas, nota-se que a sensibilidade da metodologia proposta tem um desempenho aceitável.

Através da análise dos falsos positivos por imagem, observa-se que a metodologia proposta obteve um desempenho muito bom. Pois das 3305 não-massas selecionadas na fase de segmentação, apenas 473 foram classificadas como massas, o que leva nos dá uma taxa de 0,84 falsos positivos por imagem. Deve-se notar que este índice é diretamente proporcional à especificidade. Isto nos permite ter uma idéia da especificidade não informada nos outros trabalhos.

Com a análise da área sob a curva ROC (AUC), percebe-se que a metodologia proposta obteve um bom desempenho de 0,870.

Nenhum outro trabalho utilizou o índice *overlay*, isto significa que são necessários mais estudos sobre a qualidade das técnicas de segmentação.

5. CONCLUSÃO

Mundialmente o câncer de mama apresenta alta e crescente taxa de incidência. No Brasil, ele é o que mais causa mortes entre mulheres e é o segundo maior em número casos, perdendo apenas para o câncer de pele.

Atualmente, pesquisas envolvendo técnicas de processamento de imagens têm contribuído para a detecção e diagnóstico da patologia, tornando-se assim, uma importante ferramenta de auxílio ao especialista, fornecendo uma segunda opinião, aumentando a qualidade e precisão dos exames.

Este trabalho apresentou uma metodologia CAD para a detecção de massas em imagens digitais de mamografia, utilizando Redes Neurais Celulares para segmentação das regiões de interesse. Em seguida, essas regiões de interesse têm suas características de geometria e textura extraídas, e que posteriormente serão usadas para treinar o classificador MVS que irá determinar se essa região de interesse é uma região contendo uma massa ou é uma área sadia da mama.

Os resultados apresentados no Capítulo 4 mostraram um bom desempenho da metodologia desenvolvida. A etapa de segmentação das regiões de interesse conseguiu segmentar 566 das 623 massas da amostra, o que equivale a 90,85% dos casos. A etapa de classificação das regiões segmentadas também obteve um desempenho aceitável, atingindo 84,62% de acurácia, 80,00% de sensibilidade e 85,68% de especificidade, com taxa média de falsos positivos por imagem e falsos negativos por imagem de 0,84 e 0,20 respectivamente e uma área sobre a curva ROC de 0,870. Tais resultados indicam que a combinação entre os descritores de geometria, a função K de Ripley e os Índices de Moran e Geary fornecem uma boa ferramenta para caracterizar regiões suspeitas de conterem massas. Outra importante observação é a utilização da análise envolvendo apenas a textura (com a função K de Ripley e os Índices de Moran e Geary), pois ela apresentou uma acurácia de 81,86%, sensibilidade de 81,75% e 81,86% de especificidade, com taxa média de falsos positivos por imagem e falsos negativos por imagem de 1,06 e 0,18 respectivamente e uma área sobre a curva ROC de 0,818.

Entretanto, apesar dos bons resultados obtidos, diversos aspectos da metodologia podem ser melhorados, possibilitando resultados ainda melhores. Um desses aspectos, por exemplo, é o desempenho da etapa de segmentação das regiões de interesse onde aproximadamente 9,15% das massas originais foram perdidas. Observou-se também que os objetos classificados como massa têm área menor que a área marcada pelo DDSM (índice *overlay* abaixo de 0,40 em média). Isso não chega a ser um grande problema, mas indica que massas maiores podem ser divididas em menores, o que pode influenciar na análise de textura, causando a classificação da massa como falso negativo. Um estudo extra deve ser direcionado para saber até que ponto e com que freqüência isso ocorre.

Outro problema encontrado na segmentação está no fato da grande quantidade de regiões de interesse selecionadas. Foram 566 regiões contendo massas e 3305 com áreas sadias. Essa desproporção numérica entre as duas classes acaba por influenciar o classificador, já que se dispõe de muito mais informações sobre uma classe que de outra, fato evidenciado pela necessidade do uso de pesos diferentes no treinamento do classificador MVS para poder alcançar um equilíbrio entre sensibilidade e especificidade. Uma solução pode ser alcançada através de algoritmos de otimização de parâmetros (algoritmos genéticos, sistemas *fuzzy*) para encontrar um único *template* que possa diminuir a quantidade de falsos candidatos e aumentar massas encontradas.

Além desses aspectos, diversas outras idéias surgiram ao longo do desenvolvimento deste trabalho, mas não puderam ser concluídas e incluídas no mesmo, deixando algumas possibilidades em aberto para trabalhos futuros. Entre elas estão: utilização do filtro FDOG para detecção de borda; a pesquisa de outras medidas geométricas e de textura para caracterização das massas; classificar as massas detectadas de acordo com suas naturezas malignas ou benignas; a utilização de outro classificador para comparação com os resultados obtidos pela MVS.

Por fim, o presente trabalho abre a possibilidade para utilização da metodologia aplicada para a análise de outros tipos de lesões como calcificações da mama, nódulos pulmonares, etc.

REFERÊNCIAS

ANGELO. **Sistema de Processamento de Imagens Mamográficas e Auxílio ao Diagnóstico via-Internet.** Tese de Doutorado. Pós-graduação em Engenharia Elétrica, USP, Departamento de Engenharia Elétrica. São Carlos, 2007.

ALBUQUERQUE e ALBUQUERQUE. **Processamento de Imagens: Métodos e Análises**. Centro Brasileiro de Pesquisas Físicas, 2000. Disponível em: <http://www.cbpf.br/cat/download/publicacoes/pdf/ ProcessamentoImagens.PDF>. Último acesso em 21/01/2008.

ANGELO e HAERTEL. **Investigação com respeito à aplicação dos filtros de Gabor na classificação supervisionada de imagens digitais**. Anais do X Simpósio Brasileiro de Sensoriamento Remoto, INPE, p. 1193-1200. Foz do Iguaçu, 2001.

ANSELIN. **Local Indicators of Spatial Association-LISA**. Geographical Analysis, vol. 27, nº 2, p. 91-115, 1995.

BELLOTTI, DE CARLO, TANGARO, GARGANO, MAGGIPINTO, CASTELLANO, MASSAFRA, CASCIO, FAUCI, MAGRO, RASO, LAURIA, FORNI, BAGNASCO, CERELLO, ZANON, CHERAN, LOPEZ, BOTTIGLI, MASALA, OLIVA, RETICO, FANTACCI, CATALDO, DE MITRI e DE NUNZIO. **A completely automated CAD system for mass detection in a large mammographic database**, Medical Physics, vol. 33, nº 8, p. 3066-3075, 2006.

BIRD, WALLACE e YANKASKAS. **Analysis of cancers missed at screening mammography**, Radiology, vol. 184, nº 3, p. 613-617, 1992.

BRAZ. **Identificação de Massas em Mamografias Usando Textura, Geometria e Algoritmos de Agrupamento e Classificação**. Monografia de conclusão de curso. Departamento de Ciências da Computação. Universidade Federal do Maranhão. São Luís, 2006.

BROWN e DAVIS. **Receiver operating characteristics curves and related decision measures: A tutorial**. Chemometrics and Intelligent Laboratory Systems, vol. 80, p. 24-38, 2006.

BUSHBERG, SEIBERT, LEIDHOLDT, BOONE. **The Essential Physics of Medical Imaging**. Lippincott Williams & Wilkins, 2nd Edition. Philadelphia. 2002.

CÂMARA, FUCKS, CARVALHO e MONTEIRO. **Análise Espacial de Áreas**. In: Análise Espacial de Dados Geográficos. Brasília, EMBRAPA, vol. 1, p. 157-209, 2004.

CANNY. **A Computational Approach to Edge Detection**. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, nº 6, p. 679-698, 1986.

CHAUDHURI e SARKAR. **Texture Segmentation Using Fractal Dimension**. IEEE Transactions on Pattern Analysis and Machine Intelligence, n.17, p. 72-76, 1995.

CHAVES, **Extração de Regras Fuzzy para Máquinas de Vetor de Suporte (SVM) para Classificação em Múltiplas Classes**. Tese de doutorado. Pontifícia Universidade Católica. Departamento de Engenharia Elétrica. Rio de Janeiro, 2006.

CHUA e ROSKA. **Cellular neural networks and visual computing. Foundation and applications**. Cambridge University Press, Budapest, 2004.

CHUA e YANG. **Cellular neural networks: theory**. IEEE Transactions on Systems and Circuits, vol. 35, nº 10, p. 1257–1272, 1988.

CORRÊA. **Memória associativa em redes neurais realimentadas**, Tese de mestrado. Pós-graduação em Ciências da Computação e Matemática Computacional. Universidade de São Paulo. São Paulo, 2004.

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

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

DDSM. HEATH, MICHAEL, *et al*. **The Digital Database for Screening Mammography**, in Proceedings of the Fifth International Workshop on Digital Mammography, Medical Physics Publishing, p. 212-218, 2001.

DOI. **Overview on research and development of Computer-aided diagnostic schemes**. Seminars in Ultrasound, CT and MRI, vol. 25, p. 404-410, 2004.

DUDA e HART. **Use of the Hough transformation to detect lines and curves in pictures**. Communications of the ACM, vol. 15, p. 11 – 15. New York, 1972.

EBERHART e DOBBINS. **Neural Network PC Tools, A Practical Guide**. Academic Press. San Diego, 1990.

EBERT. **Texturing and Modeling: A Procedural Approach**. Academic Press, Cambridge, 1994.

EFROYMSON. **Multiple regression analysis**. In Ralston, A. and Wilf, HS, editors, Mathematical Methods for Digital Computers. Wiley. 1960.

ELTONSY, TOURASSI e ELMAGHRABY. **A concentric morphology model for the detection of masses in mammography**, IEEE Transactions on Medical Imaging, vol. 26, nº. 6, p. 880–889. Davis, 2007.

FENTON, TAPLIN, CARNEY, ABRAHAM, SICKLES, D'ORSI, BERNS, CUTTER, HENDRICK, BARLOW e ELMORE. **Influence of Computer-Aided Detection on Performance of Screening Mammography**. Breast Diseases: A Year Book Quarterly, vol. 18, 2007.

FONSECA. **Processamento e análise de imagens aplicados à caracterização automática de materiais**. Dissertação de Mestrado, Pós-graduação em Ciências da Engenharia Metalúrgica. Departamento de Ciência de Materiais e Metalurgia, PUC, Rio de Janeiro, 2001.

FREER, ULISSEY. **Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center**. Radiology, vol. 220, p. 781-786, 2001.

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

GONZALEZ e WOODS. **Digital Image Processing**. 3ª ed. Prentice Hall. Mazidi, 2007.

GRIFFITH. **Spatial autocorrelation**. Association of American Geographers. Washington, 1987.

HAN e KAMBER. **Data Mining – Concepts and Techniques**, 2ª ed., Ed. Morgan Kaufmann Publishers, San Francisco, 2006.

HASSANIEN. **Fuzzy rough sets hybrid scheme for breast cancer detection**. Image and Vision Computing. vol. 25, p. 172-183, Butterworth-Heinemann. Newton 2007.

HAYKIN e ENGEL. **Redes Neurais: Princípios e Prática**. 2ª ed., Bookman. Hamilton, 2008.

HEATH, BOWYER, KOPANS, KEGELMEYER, MOORE, CHANG e KUMARAN. **Current Status of the Digital Database for Screening Mammography**. Digital Mammography. Proceedings of the Fourth International Workshop on Digital Mammography, p. 457–460, Kluwer Academic Publishers. Netherlands, 1998.

HOUGH. **Machine Analysis of Bubble Chamber Pictures**. Proceedings of International Conference on High Energy Accelerators and Instrumentation. Geneva, 1959.

INCA - **Instituto Nacional do Câncer**. Disponível em: <http://www.inca.gov.br/index.asp>. Última visita em 13/06/2009.

IREANEUS e THAMARAI. **Digital mammogram segmentation and tumor detection using artificial neural networks**. International Journal of Soft Computing, p. 112-119, 2008.

JAIN, MURTY e FLYNN. **Data Clustering: A Review**. ACM Computing Surveys, vol. 31, nº 3, p. 264-323. New York, 1999.

KARSSEMEIJER, OTTEN, VERBEEK, GROENEWOUD, KONING, HENDRIKS e HOLLAND. **Computer-aided detection versus independent double reading of masses on mammograms**. Radiology, vol. 227, p. 192-200, 2003.

KERLIKOWSKE, CARNEY, GELLER, MANDELSON, TAPLIN, MALVIN, ERNSTER, URBAN, CUTTER, ROSENBERG e BALLARD-BARBASH. **Performance of screening mammography among women with and without a first-degree relative with breast cancer**. Annals of Internal Medicine, vol. 133, nº 11, p. 855–863, 2000.

KIRALJ e FERREIRA. **Basic Validation Procedures for Regression Models in QSAR and QSPR Studies: Theory and Application**. Journal of the Brazilian Chemical Society, vol. 20, nº 4, p. 770-787, 2009.

KOLLER e SAHAMI. **Toward optimal feature selection**. International Conference on Machine Learning, p. 284–292. Bari, 1996.

KOM, TIEDEU e KOM. **Automated detection of masses in mammograms by local adaptive thresholding**, Computers in Biology and Medicine, vol. 37, nº 1, p. 37-48, 2007.

KOPANS, CHAN, WEI, SAHINER, RAFFERTY, WU, ROUBIDOUX, MOORE, HADJIISKI e HELVIE. **Computer-aided Detection System for Breast Masses on Digital Tomosynthesis Mammograms: Preliminary Experience**. Radiology, vol. 237, p. 1075 – 1080, 2005.

KOPANS. **Imagem da Mama**. 2ª ed. MEDSI Editora Médica e Científica Ltda. Rio de Janeiro, 2000.

LANCASTER e DOWNES. **Spatial Point Pattern Analysis of Available and Exploited Resources**. Ecography, vol. 27, nº 1, p. 94-102. Copenhagen, 2004.

LANGLEY e IBA. **Average-case analysis of a nearest neighbor algorithm**. International Joint Conference on Artificial Intelligence, p. 889-894, Chambery, 1993.

LI. **Markov Random Field, Modeling in Computer Vision**. 2ª ed. Springer-Verlag. New York, 2001.

MAMAINFO. MAMAinfo é um portal de informações atualizadas sobre câncer de mama que desenvolve programas educacionais e sociais, colaborando na luta contra o câncer. Disponível em <**http://www.mamainfo.org.br/**>. Último acesso em 01/06/2009.

MANDAL, IDRIS e ANCHANATHAN. **A critical evaluation of image and video indexing techniques in the compressed domain**. Image and Vision Computing Journal, vol. 17 p. 513–529, 1999.

MARTINS, SILVA, PAIVA e GATTASS. **Detection of Breast Masses in Mammogram Images Using Growing Neural Gas Algorithm and Ripley's K Function**. Journal of Signal Processing Systems, vol. 55, p. 77-90, Kluwer Academic Publishers. Hingham, 2009.

MAZUROWSKI, HABAS, ZURADA, LO, BAKER e TOURASS. **Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance**. Neural Networks, ed. 21, p. 427-436, 2008.

MEDRONHO, BLOCH, LUIZ e WERNECK. **Epidemiologia**, Atheneu, São Paulo, 2004.

MONTORO. **Prevenção e Detecção do Câncer de Mama**. McGraw-Hill do Brasil, p.147-163. São Paulo, 1979.

NASCIMENTO e RAMOS. **Combinando duas visões mamográficas em extração de características com Ridgelet**. Anais do XI Congresso Brasileiro de Informática em Saúde. Campos do Jordão, 2008.

NUNES-1. **Detecção de massas em imagens mamográficas usando índice de diversidade de Simpson e máquina de vetores de suporte**. Tese de mestrado, Departamento de Engenharia Elétrica, Universidade Federal do Maranhão. São Luís, 2009.

NUNES-2. **Investigações em Processamento de Imagens Mamográficas para Auxílio ao Diagnóstico de Mamas Densas**. Tese de Doutorado em Ciências. Pós-graduação em Física Computacional, Universidade de São Paulo, São Carlos, 2001.

PADWAL, **Elements of breasting imaging basics**. Disponível em <http://www.gehealthcare.com/usen/ultrasound/education/products/ cme_breast.htm>. Último acesso em 31 ago. 2009.

PAIVA, RODRÍGUEZ e CORREIA. **Métodos Computacionais para Analisar Padrões de Pontos Espaciais**. INPE, 1999. Disponível em <http://www.dpi.inpe.br/geopro/trabalhos/gisbrasil99/ estat_pontos>. Acesso em 31 ago. 2009.

PAL e PAL. **A Review on Image Segmentation Techniques**. Pattern Recognition, Vol. 26, p. 1277–1294, 1993.

RANGAYYAN, FARAMAWY, DESAUTELS e ALIM. **Measures of Acutance and Shape for Classification of Breast Tumors**. IEEE Transactions on Medical Imaging, vol. 16, p. 799-810. Davis, 1997

RIPLEY. **Modelling spatial patterns**. Journal of the Royal Statistical Society, vol. 39, p. 172–212, 1977.

SCHOUTEN. **Image Processing**. Radboud University, Department of Computer Science. Disponível em <http://www.cs.ru.nl/~ths/rt2/col/h9/9gebiedENG.html>. Acesso em 23 out. 2007.

SERHAT, ONUR e YILMAZ. **Mammographic Mass Detection Using a Mass Template**. Korean Journal of Radiology, vol. 6, p. 221–228. 2005.

SOARES, REZENDE e FORTES. **Calibração multivariada com seleção de variáveis em amostras de biodiesel adulteradas com óleo de soja cru utilizando dados de espectroscopia de infravermelho com transformada de Fourier**. Anais II Congresso da Rede Brasileira de Tecnologia de Biodiesel. Brasília, Supernova Design, vol 2. p. 95-96, 2007.

SOUKUP e DAVIDSON. **Visual Datamining – Techniques and Tools for Data Visualization and Mining**, Ed. Wiley Publishing, Incorporated. 2002.

SOUSA, SILVA e PAIVA. **Lung Structures Classification Using 3D Geometric Measurements and SVM**. 12th Iberoamerican Congress on Pattern Recognition, Valparaiso, vol. 4756, p. 783-792, Springer-Verlag. Berlin, 2007.

SOVIERZOSKI, ARGOUD E AZEVEDO. **Avaliação do classificador neural binário com análise ROC**. 21º Congresso Brasileiro de Engenharia Biomédica, p. 989-992. Salvador, 2008.

STRICKLAND. **Image-processing techniques for tumor detection**, Marcel Dekker, p. 238-239. New York, 2002.

SUCKLING, PARKER, DANCE, ASTLEY, HUTT e BOGGIS. **The mammographic images analysis society digital mammogram database**. Exerpta Medica. International Congress Series, vol. 1069, p. 375-378. 1994.

THURFJELL, LENERVALL e TAUBE. **Benefit of independent double reading in a population-based mammography screening program.** Radiology, vol.191, p. 241-244, 1994.

TIMP, VARELA e KARSSEMEIJER. **Temporal change analysis for characterization of mass lesions in mammography**. IEEE transactions on medical imaging, vol. 26, p. 945-953, 2007.

TÓTH, TAKÁCS e PATAKI. **Mass detection in mammograms combining two methods**. 3rd European Medical & Biological Engineering Conference, Praga, 2005.

TUCERYAN e JAIN. **Texture Analysis. The Handbook of Pattern Recognition and Computer Vision**. 2ª ed., p. 207-248, World Scientific Publishing. 1998.

VALE e POZ. **O Processo de Detecção de Bordas de Canny: Fundamentos, Algoritmos e Avaliação Experimental**. Simpósio Brasileiro de Geomática, Presidente Prudente. Anais do Simpósio Brasileiro de Geomática, vol. 1. p. 292-303, 2002.

VALLE. **Câncer de Mama Locorregional Avançado**. MEDSI. Rio de Janeiro, 1999.

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

WANG e KARAYIANNIS. **Detection of microcalcifications in digital mammograms using wavelets**, IEEE Transactions in Medical Imaging, vol. 17, nº 4, p. 498–509, 1998.

WERKEMA e AGUIAR. **Análise de Regressão: como entender o relacionamento entre as variáveis de um processo**. Série Ferramentas

da Qualidade, Escola de Engenharia da UFMG, vol. 7. Belo Horizonte, 1996.

WIRTH. **Shape Analysis and Measurement**. Lecture Notes on Image Processing, Universidade de Guelph, 2001. Disponível em: <http://www.uoguelph.ca/ ~mwirth/cis6320/lec10notes.pdf>. Último acesso em 16/11/2007.

WITTEN e FRANK. **Data Mining: Practical machine learning tools and techniques**, 2ª ed., Morgan Kaufmann, 2005.

XING. **Understanding and Using Microarray Analysis Techniques: A Practical Guide**, Kluwer Academic Publishers, London, 2003.

YING, XINBO e JIE. **A feature analysis approach to mass detection in mammography based on RF-SVM**. Image Processing, vol. 5, p. 9–12, IEEE International Conference. San Antonio, 2007.

ZARANDY, ROSKA, LISZKA, HEGYESI, KEK e REKECZKY. **Design of Analogic CNN Algorithms for Mammogram Analysis**. Third IEEE International Workshop on Cellular Neural Networks and their Applications, p. 255-260. Rome, 1994.