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

Detecção de Regiões de Massas em Mamografias usando índices de Diversidade, Geoestatística e Geometria Côncava

---

Geraldo Braz Junior

São Luís
2014

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

Detecção de Regiões de Massas em Mamografias usando índices de Diversidade, Geoestatística e Geometria Côncava

---

Geraldo Braz Junior

Tese apresentada ao Programa de
Pós-Graduação em Engenharia de Eletricidade da UFMA
como parte dos requisitos necessários para obtenção do
grau de Doutor em Engenharia Elétrica.

Orientadores: **Prof. Dr. Anselmo Cardoso de Paiva**
Prof. Dr. Aristófanes Corrêa Silva

São Luís
2014

Braz Junior, Geraldo.

Detecção de regiões de massas em mamografias usando índices de diversidade, geoestatística e geometria côncava/ Geraldo Braz Junior – São Luís, 2014.

118 f.

Orientador: Anselmo Cardoso de Paiva

Co- orientador: Aristófanes Côrrea Silva

. Tese (Doutorado) – Universidade Federal do Maranhão, Programa de Pós-Graduação em Engenharia de Eletricidade, 2014.

1. Mamografia – Câncer de mama. 2. Análise de diversidade. 3. Geoestatística. 4. Geometria côncava. I. Título.

CDU 004.383.5:618.19-006

DETECÇÃO DE REGIÕES DE MASSAS EM MAMOGRAFIAS USANDO ÍNDICES DE DIVERSIDADE, GEOESTATÍSTICA E GEOMETRIA CÔNCAVA

Geraldo Braz Júnior

Tese aprovada em 10 de março de 2014

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

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

Prof. Hélio Côrtes Vieira Lopes, Dr.
(Membro da Banca Examinadora)

Prof. Leandro Augusto Frata Fernandes, Dr.
(Membro da Banca Examinadora)

Prof. João Viana da Fonseca Neto, Dr.
(Membro da Banca Examinadora)

Prof. Zair Abdelouahab, Ph.D.
(Membro da Banca Examinadora)

"Coloque a lealdade e a confiança acima de qualquer coisa;
não te alies aos moralmente inferiores;
não receies corrigir teus erros."

Confucio

Agradecimentos

À Deus;
Pelo apoio incondicional de minha família, em especial de Vandecia;
Pela dedicação, conselhos e direcionamento de meus orientadores e amigos Anselmo e Ari;
Pelo apoio a mim oferecido, por parte de meus amigos, durante os momentos não tão felizes dessa trajetória;
A todos aqueles que direta ou indiretamente participaram deste trabalho, contribuindo para o sucesso do mesmo.

Resumo

O câncer de mama se configura como um problema de saúde mundial, que afeta principalmente a população feminina. É conhecido que a detecção precoce aumenta as chances de um tratamento efetivo, melhorando o prognóstico da doença. Com este objetivo, ferramentas computacionais têm sido propostas com a finalidade de auxiliar o especialista na interpretação do exame de mamografia, provendo funcionalidades de detecção e diagnóstico de lesões. Todavia, continua sendo um grande desafio detectar a lesão com alta taxa de sensibilidade, e garantir ao mesmo tempo que um número reduzido de falso positivos sejam gerados. Para tanto, metodologias que abordam extração de características texturais, probabilísticas ou baseada em modelo têm sido propostas para este fim. A pesquisa que remete este trabalho tem como objetivo principal a proposição de uma metodologia eficiente de detecção de regiões de massas em mamografias digitalizadas. A tarefa de detecção envolve aspectos de visão computacional relacionados a necessidade de encontrar regiões suspeitas e descrevê-las de maneira discriminatória. Esta pesquisa avalia a extração de características usando as abordagens de análise de diversidade, geoestatística e geométrica para a classificação das regiões suspeitas detectadas usando a Máquina de Vetores de Suporte como classificador. Os resultados encontrados são promissores ao obterem alta sensibilidade e baixa taxa média de falso positivos quando usando geometria côncava para extrair características.

Palavras-Chave: Câncer de Mama, Análise de Diversidade, Geoestatística, Geometria Côncava, Reconhecimento de Padrões.

Abstract

Breast cancer is configured as a global health problem that affects mainly the female population. It is known that early detection increases the chances of an effective treatment and improves the prognosis of the disease. With this goal, computational tools have been proposed in order to assist the physician in the interpretation of mammography features providing detection and diagnosis of lesions. The challenge is to detect any lesion with high sensitivity rate while maintaining a small number of false positives. The main objective of this research is the development of an efficient methodology for mass detection in digitized mammograms. The detection task involves aspects of computer vision like find suspicious areas and describe them in a discriminatory way. This research evaluates the approaches of feature extraction using diversity analysis, geostatistics and concave geometry for the classification of previously identified suspicious regions using Support Vector Machine as a classifier technique. The results are promising and reaches a high sensitivity rate jointly with a low mean rate of false positives per image when using concave geometry as feature extraction approach.

Keywords: Breast Cancer, Diversity Analysis, Spatial Statistics, Concave Geometry, Pattern Recognition.

Lista de Tabelas

Lista de Figuras

Sumário

Capítulo 1

Introdução

De acordo com o Instituto Nacional de Câncer (INCA), o termo câncer é utilizado genericamente para designar um conjunto de mais de 100 doenças, incluindo tumores malignos de diferentes localizações. Nas últimas décadas, o número de incidência tem aumentado e a estimativa da Organização Mundial de Saúde (OMS) para o ano de 2030 é de 27 milhões de casos de câncer, com 17 milhões de mortes (BOYLE *et al.*, 2008). Esta também é a segunda causa de morte entre a população mundial (WHO, 2012).

As estatísticas consolidadas do ano de 2008 por Boyle *et al.* (2008) e acompanhadas no projeto GLOBOCAN demonstram que o tipo de câncer mais frequente após o melanoma é o câncer de pulmão (1,6 milhões de casos novos), seguido pelo de mama (1,4 milhões) e cólon e reto (1,35 milhões). Devido ao mau prognóstico, o câncer de pulmão foi a principal causa de morte (1,31 milhões), seguido pelo câncer de estômago (750 mil óbitos) e pelo câncer de fígado (720 mil óbitos).

Cerca de 30% das mortes por câncer são devidas aos cinco principais riscos comportamentais e alimentares, sendo eles: índice de massa corporal elevado; baixa ingestão de frutas, legumes e verduras; falta de atividade física; tabagismo e uso de álcool (BOYLE *et al.*, 2008).

No Brasil, estimativas para o ano 2014 (INCA, 2014) apontam que 576 mil novos casos de câncer serão registrados (contra 518.510 estimados para 2013), onde 181.550 casos são de câncer de pele não melanoma. Os demais casos são distribuídos 203.930 casos para o sexo masculino e 190.520 no grupo

feminino. Para o grupo feminino, o câncer de mama será o predominante, representando 57.120 novos casos, 20,8% de todos os novos casos de câncer no grupo feminino (INCA, 2014). Ainda segundo o INCA (2014), a região Nordeste tem uma estimativa para o ano de 2014 de 10.490 novos casos de câncer de mama, seguindo a média de incidência brasileira.

Visando promover a saúde pública de qualidade, o governo brasileiro criou a Lei 11.664/2008 que entrou em vigor em 29 de abril de 2009 e que dispõe sobre a efetivação de ações de saúde que assegurem a prevenção, a detecção, o tratamento e o seguimento dos cânceres do colo uterino e de mama, no âmbito do Sistema Único de Saúde − SUS. Além disso, diversas campanhas foram feitas a fim de conscientizar as mulheres, principalmente as que estão dentro do grupo de risco.

Uma das maneiras para detectar os tumores não apalpáveis que causam câncer de mama é realizar uma mamografia das mamas conforme programas de rastreamento. A mamografia é atualmente a melhor técnica de detecção precoce de lesões não apalpáveis na mama com altas chances de ser um câncer curável. A partir do início da utilização da mamografia foi observada uma redução da taxa de mortalidade associada a essa patologia (SOCIETY, 2013). Entretanto, a sensibilidade desse exame pode variar bastante, em decorrência de fatores como qualidade do exame, experiência do especialista ou idade da paciente, resultando em falhas nos laudos emitidos por radiologistas que variam entre 10% a 30% (BIRD *et al.*, 1992).

A mamografia, nos programas de rastreamento, é proposta apenas para mulheres acima dos quarenta anos de idade. Este exame é bastante efetivo para tal faixa etária, o que fornece uma boa relação custo/benefício. Mulheres abaixo dessa idade são acompanhadas através de exames de ultrassonografia ou, se pertencerem ao grupo de risco, por exames de ressonância magnética, quando disponível. Os tipos mais comuns de anormalidades visíveis em imagens de mamografia são: calcificações (benignas e malignas); massas circulares e bem definidas; massas espiculadas; massas mal definidas e distorção de arquitetura (HEATH *et al.*, 1998).

Uma massa é um aglomerado de células que se unem de forma mais densa em relação ao tecido que a envolve. Este aglomerado pode ser causado por câncer de mama, assim como também por condições benignas. Algumas características das massas são determinantes para estabelecer suas probabilidades de malignidade,

como por exemplo tamanho, forma e disposição de suas margens.

Já as calcificações são depósitos de cálcio que aparecem como pontos brancos na imagem da mamografia. São de dois tipos: microcalcificações e macrocalcificações. As microcalcificações são depósitos pequenos e indicam, dependendo de sua forma, uma possível presença cancerígena. As macrocalcificações são grandes depósitos de cálcio e normalmente estão associadas com condições benignas, causadas, por exemplo, por inflamações ou envelhecimento das artérias.

Nesse contexto, técnicas computacionais de processamento de imagens e reconhecimento de padrões têm adquirido importância para o diagnóstico e auxílio na intervenção médica. O tempo gasto para trabalhar com essas imagens, a subjetividade dos atributos extraídos e a necessidade contínua de investigação para o progresso na área, têm contribuído para o surgimento de novas metodologias de processamento e análise das imagens médicas que melhoram a qualidade do diagnóstico médico.

O processamento de imagens na medicina representa um conjunto de técnicas computacionais, que aplicadas, podem prover auxílio ao diagnóstico, planejamento de tratamentos, simulação de cirurgias, compressão de imagens em bancos de exames, recuperação de exames por conteúdo de imagens, auxílio à pesquisa em medicina, educação médica, dentre outros.

O objetivo do uso do processamento digital de imagens consiste em melhorar o aspecto visual de certas feições estruturais para o analista humano e fornecer subsídios para a sua interpretação, inclusive gerando produtos que possam ser posteriormente submetidos a outros processamentos. A evolução da tecnologia de computação digital, bem como o desenvolvimento de novos algoritmos para lidar com sinais bidimensionais está permitindo uma gama de aplicações cada vez maior.

Sistemas de detecção e diagnóstico auxiliado por computador (respectivamente CAD - *Computer-Aided Detection* e CADx - *Computer- Aided Diagnosis*) têm sido propostos com o objetivo de auxiliar o radiologista, indicando áreas suspeitas, bem como anormalidades mascaradas. Esses sistemas têm sido desenvolvidos por vários grupos de pesquisa, visando auxiliar na detecção e diagnóstico precoce do câncer de mama (MEERSMAN *et al.*, 1998). Estudos mostram que o índice de

detecção da presença de câncer de mama poderia ser aumentado de 5% a 15% se ferramentas CAD fossem utilizadas (FREER; ULISSEY, 2001).

As ferramentas de detecção auxiliam os especialistas na interpretação do exame e no planejamento de procedimentos invasivos e as ferramentas de diagnóstico na decisão a respeito da realização de certos procedimentos, os quais, tomados em um espaço de tempo curto, podem ser fundamentais em um tratamento adequado e com grandes chances de sucesso. Juntas, as ferramentas de detecção e diagnóstico constituem uma importante ferramenta de auxílio ao especialista para promover o desenvolvimento de tratamentos mais adequados aos pacientes.

1.1 Definição do Problema

Embora as ferramentas de detecção representem uma melhora na sensibilidade do exame, características da imagem de mamografia, como a sobreposição de tecidos, a limitação de representação de textura e a subjetividade na análise da imagem, implicam na necessidade de construção de técnicas eficientes em termos de sensibilidade ao mesmo tempo que geram uma quantidade reduzida de erros, tratados como falso positivos.

Detalhadamente, o problema tratado nessa tese consiste em:

- Criar uma técnica de detecção de regiões suspeitas que tenha altas taxas de sensibilidade (superior ao padrão gold, estimado em média a 85%) ao mesmo tempo que tenha um baixo volume de erros, considerando que existe uma imensa heterogeneidade entre as imagens de mamografia;

- Adaptar técnicas de processamento de imagens para o tratamento genérico das imagens de mamografia, nas área de:

  - Melhoramento e segmentação das feições internas e
  - Descrição objetiva e discriminatória de objetos identificados nas feições internas.

- Determinar um padrão distinguível entre regiões suspeitas e falso positivos para promover a minimização de erros e

- Adaptar técnicas de aprendizado de máquina para o tratamento genérico dos padrões.

1.2 Trabalhos Relacionados

As imagens de mamografia representam as estruturas internas da mama com pouca informação e com sobreposição. Normalmente, as massas são estruturas muito semelhantes a outras ao seu redor ou se apresentam na imagem de maneira oclusa. Isso torna as metodologias de detecção de massas muito específicas e pouco parametrizáveis. Também faz com que vários grupos de pesquisa tenham estudado o tema em busca de uma metodologia eficiente de detecção e portanto propondo a mesma sob diferentes óticas. Em geral, essas metodologias são compostas de duas etapas claras: segmentação da massa e redução de falso positivos.

Na maioria das metodologias, a segmentação é realizada somente sobre uma imagem. Todavia, uma nova classe de estudos tem considerado o uso das incidências MLO[1] e CC[2] conjuntamente (intitulado ipsilateral) para detectar massas pelas diferenças entre as mesmas (ENGELAND; KARSSEMEIJER, 2007) (QIAN *et al.*, 2007) (WEI *et al.*, 2011b). Um princípio similar é aplicado a metodologias que usam a visão bilateral (WU *et al.*, 2007) (KE *et al.*, 2010) (TZIKOPOULOS *et al.*, 2011) (WANG *et al.*, 2012). Se uma lesão é percebida em apenas uma visão, pode se tratar na verdade de uma distorção arquitetural (ACR, 2003a).

A detecção de massas é uma tarefa complexa pela grande variedade de possibilidades que o algoritmo deve lidar. Métodos baseados em densidade como *MeanShift* (SAHBA; VENETSANOPOULOS, 2010a), (SAHBA; VENETSANOPOULOS, 2010b), (TERADA *et al.*, 2010) tem recentemente sido aplicados com sucesso para seleção de regiões densas em mamografias. Alguns trabalhos realizam uma pré-classificação de densidade da imagem (OLIVER *et al.*, 2010) (TZIKOPOULOS *et al.*, 2011) como forma parametrizável de ajuste de parâmetros. A suposição é que massas correspondem a regiões de grande intensidade e se situam em contraste ao restante das glândulas mamárias.

[1] Médio-Lateral Oblíqua
[2] Crânio-Caudal

Métodos geométricos, como *Level Sets* (YUAN *et al.*, 2007), Múltiplas Camadas Concêntricas (ELTONSY *et al.*, 2007), *Morphological Component Analysis* (GAO *et al.*, 2010), *Isocontour Map* (HONG; SOHN, 2010), *Circular Gaussian Filter* (ZHENG, 2010) e Contornos Ativos (RAHMATI *et al.*, 2012), também se mostram efetivos por apresentar resultados superiores a 90% para detecção de massas em imagens de mamografias levando em consideração o aspecto morfológico de massas e outros tecidos visíveis na imagem. Outras metodologias utilizam também com sucesso a detecção baseada em informações de textura (KOM *et al.*, 2007) e (MAZUROWSKI *et al.*, 2011).

Para a etapa de redução de falso positivos, normalmente características de textura e geometria são extraídas como prática de alicerçar o reconhecimento de padrões. *Local Binary Patterns* ((LLADÓ *et al.*, 2009) (LIU *et al.*, 2011)), *Countourlet* (MOAYEDI *et al.*, 2010), *Ridgelet* (RAMOS *et al.*, 2012), *Generalized Moment Patterns* (DEEPAK *et al.*, 2012), *Adaptive Median Filtering* (BASHEER; MOHAMMED, 2013), *GLCM-Optical Density Features* (TAI *et al.*, 2013) e *Gray-Scale Invariant Ranklet Texture Features* (MASOTTI *et al.*, 2009) são alguns exemplos de metodologias de extração de textura usadas em etapas de redução de falso positivos. Uma característica comum é que realizam a tarefa usando informação estatística mútua baseada em fatores de localização, ocorrência conjunta ou multirresolução.

Um grande desafio das metodologias automáticas de detecção está na redução da quantidade de regiões suspeitas que são assinaladas. Essa quantidade normalmente é grande devido ao fato da massa ser relativamente semelhante a outras estruturas que apresentam intensidade e forma similar após o processo de detecção.

Caso o número de regiões assinaladas como suspeitas seja baixo, é necessário verificar se dentro deste conjunto também estão as massas que deveriam ser detectadas. A medida que resume essa análise é a sensibilidade. Logo, um método de detecção com baixa sensibilidade pode levar a interpretação da inexistência de uma lesão (seja a mesma benigna ou maligna) e portanto gerar um resultado ainda mais maléfico do que gerar muitas regiões que internamente englobem o conjunto de todas as regiões verdadeiramente massas.

A Tabela 1.1 apresenta um resumo do desempenho encontrado pelas

metodologias citadas nessa seção. A informação de base de imagens utilizada é de difícil comparação pelo fato dos autores utilizarem subconjuntos destas bases ou mesmo bases privadas. Os trabalhos são comparados em termos de acurácia (Acc), sensibilidade (S), especificidade (Sp), área sob a curva ROC[3] (ROC), área sob a curva FROC[4] (AFROC) e número médio de falso positivos por imagem (Fp/i).

Tabela 1.1: Comparação do desempenho das metodologias de detecção e redução de falso positivos apresentadas na seção de trabalhos relacionados.

Trabalho	Base	Acc	S	Sp	ROC	AFROC	FP/i
(ENGELAND; KARSSEMEIJER, 2007)	Privada	–	61,00	–	–	–	0,1
(QIAN *et al.*, 2007)	Privada	–	91,00	–	–	–	0,875
(ELTONSY *et al.*, 2007)	DDSM	–	81,00	–	–	–	0,6
(KOM *et al.*, 2007)	Privada	–	95,91	–	–	–	0,033
(WU *et al.*, 2007)	Privada	–	85,00	–	–	–	1,15
(LLADÓ *et al.*, 2009)	DDSM	93,00	–	–	0,94	–	–
(MASOTTI *et al.*, 2009)	DDSM	–	70,00	–	–	–	0,92
(MOAYEDI *et al.*, 2010)	MIAS	91,52	–	–	–	–	–
(OLIVER *et al.*, 2010)	Privada	–	86,70	–	–	0,946	2
(GAO *et al.*, 2010)	DDSM	–	95,3	–	–	–	2,83
(HONG; SOHN, 2010)	DDSM	–	90,00	–	–	–	2,3
(ZHENG, 2010)	Privada	–	93,00	–	–	–	1,19
(MAZUROWSKI *et al.*, 2011)	Privada	71,90	61,50	–	–	–	2
(KE *et al.*, 2010)	MIAS	–	85,11	–	–	–	1,44
(SAHBA; VENETSANOPOULOS, 2010a)	MIAS	–	88,00	–	0,86	–	2,1
(SAHBA; VENETSANOPOULOS, 2010b)	MIAS	–	90,00	–	0,88	–	1,9
(TERADA *et al.*, 2010)	Privada	–	81,2	–	–	–	5,0
(WEI *et al.*, 2011a)	Privada	–	87,00	–	–	–	1
(LIU *et al.*, 2011)	DDSM	–	76,8	–	–	–	1,36
(DEEPAK *et al.*, 2012)	MIAS	98,90	100,00	97,00	0,98	–	–
(WANG *et al.*, 2012)	Privada	–	84,00	–	–	–	0,69
(RAHMATI *et al.*, 2012)	DDSM	–	86,85	–	–	–	–
(TAI *et al.*, 2013)	DDSM	–	90,3	–	–	–	4,8

Através da análise dos trabalhos relacionados, podemos concluir que metodologias de detecção de massas enfrentam um grande obstáculo quanto à

[3]A curva ROC (*Receiver Operating Characteristic*) (OBUCHOWSKI, 2005) é uma forma de representar a relação, normalmente antagônica, entre a sensibilidade e a especificidade de um teste diagnóstico quantitativo ao longo de valores contínuos de ponto de corte.

[4]*Free Receiver Operating Characteristic* (GUR *et al.*, 2009) representa para metodologias de detecção a relação de precisão entre sensibilidade e taxa média de falso positivos por imagens.

necessidade de se adaptar a diferentes tipos de imagens e ainda a características de textura. Essa conclusão é alicerçada na análise da Tabela 1.1 onde metodologias que utilizam bases de imagem pública (DDSM[5] ou MIAS[6]) possuem alta taxa de sensibilidade normalmente associada a altas taxas médias de falso positivo por imagem. A situação inversa acontece, quando a sensibilidade é baixa.

Evidentemente que a sensibilidade deve ser priorizada em relação a falso positivos. Esse fator proporciona uma imprecisão que gera normalmente um grande número de falso positivos, motivando outras metodologias especializadas na redução destes. Verificamos a oportunidade para o desenvolvimento de uma metodologia para utilização conjunta de características de forma e textura como aspecto para auxílio à detecção de massas com precisão.

1.3 Solução e Objetivos

O objetivo geral desta tese é desenvolver uma metodologia para detecção e reconhecimento de regiões de massa em mamografias digitalizadas de maneira automatizada, que proporcione alta sensibilidade e resulte em baixa taxa média de falso positivos por imagem.

A solução proposta por esta tese para atingir o objetivo geral se dá através da utilização e adaptação de: algoritmos de segmentacão *MeanShift* (CHENG, 1995) e *Fast Scanning* (DING *et al.*, 2009), análise espacial de textura usando índices de diversidade (BRAZ JUNIOR *et al.*, 2013) e índices geoestatísticos (BRAZ JUNIOR *et al.*, 2009), análise geométrica usando geometria côncava (MUCKE, 1994), e reconhecimento de padrões usando Máquina de Vetores de Suporte (VAPNIK, 1998).

De maneira específica, pretendemos:

- Estudar, adaptar e aplicar algoritmos de segmentação e agrupamento baseados em crescimento de regiões (*Fast Scanning*) e densidade (*MeanShift*) para reduzir a subjetividade da imagem de mamografia em estruturas semelhantes a massas;

[5]DDSM (*Digital Database for Screening Mamography*) (HEATH *et al.*, 1998).

[6]*The Mammographic Image Analysis Society Digital Mammogram Database* (SUCKLING *et al.*, 1994).

- Estudar, adaptar e aplicar índices de diversidade e análise espacial para descrição de textura de regiões suspeitas extraídas da mamografia com finalidade de reconhecimento de regiões de massa;
- Estudar, adaptar e aplicar o uso de geometria côncava para descrição de forma para melhorar a eficiência de métodos de descrição baseados somente em textura e auxiliar no reconhecimento de regiões de massa; e
- Avaliar a metodologia proposta através de experimentos, usando bases públicas de mamografias digitalizadas.

1.4 Contribuições

Destacamos como principais contribuições desta tese:

- Construção de mecanismos adaptáveis de controle de segmentação baseada em densidade usando *MeanShift* e *Fast Scanning*;
- Construção de técnicas de zoneamento para o realce da extração de características em índices de diversidade e geoestatísticos;
- Extensão de índices de diversidade através da adição de informação espacial na distribuição de espécies com a finalidade de análise de textura de regiões extraídas de mamografias;
- Extensão de índices geoestatísticos para caráter local com finalidade de análise de textura de regiões extraídas de mamografias; e
- Construção de mecanismos de geração de características texturais e geométricas baseadas em contornos côncavos de regiões extraídas de mamografias através da utilização de *Alpha-Shapes* e adaptação de conceitos de diversidade.

1.5 Organização da Tese

O restante desta tese está organizado em mais quatro capítulos:

- O Capítulo 2 trata da fundamentação teórica necessária para construção desta pesquisa. São abordados os temas referentes ao câncer de mama, imagens radiográficas da mama, técnicas de pré-processamento de imagens, segmentação por *MeanShift* e *Fast Scanning*, análise de diversidade, análise espacial, análise geométrica e geometria côncava, reconhecimento de padrões e máquinas de vetores de suporte;

- O Capítulo 3 trata de todas as etapas da metodologia objeto desta tese, incluindo as adaptações realizados nos algoritmos de detecção e extração de características para cumprirem o objetivo da tese; as etapas de redução de falso positivos usando extração de características de textura baseada em índices de diversidade, geoestatísticos e geometria côncava;

- O Capítulo 4 apresenta os resultados obtidos utilizando a metodologia proposta sobre as bases públicas de mamografias DDSM e MIAS; estudo de casos de sucesso e falhas; e

- O Capítulo 5 apresenta a discussão dos resultados obtidos, as contribuições atingidas por este trabalho e ainda a proposta de trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo é apresentado o referencial teórico utilizado para elaboração da metodologia proposta. As seções seguintes exploram a imagem de mamografia e suas caraterísticas, processamento de imagens digitais e as subáreas de realce de imagens, segmentação, extração de características de textura baseada em estatística espacial, diversidade de espécies, extração de características baseada em geométrica côncava e reconhecimento de padrões.

2.1 Fisiologia, Imagem e Patologia da Mama

O câncer de mama é caracterizado pela disfunção ou distúrbio na reprodução ou mortalidade das células que compõe o tecido mamário. Estes distúrbios podem ocasionar o aparecimento de tumores, exemplificados pela Figura 2.1, dentre outros achados.

A relação da estrutura anatômica das glândulas mamárias em relação as imagens radiológicas adquiridas dela é de suma importância para o entendimento da grande variedade e diferenças sutis que podem ou não representar um achado.

Fisiologia e Câncer de Mama

Um aspecto importante acerca da composição fisiológica das glândulas mamárias é que não existe duas pessoas com mamas estruturalmente semelhantes e que essa estrutura individualmente sofre alterações ao longo do tempo, causadas

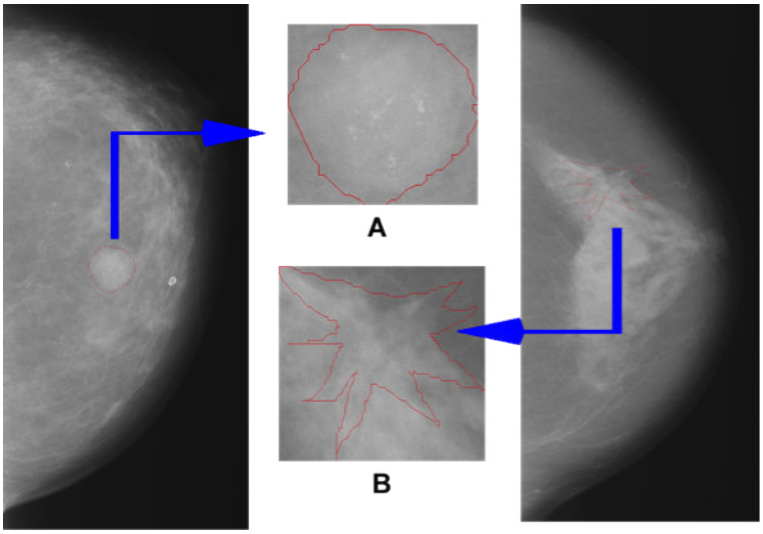


Figura 2.1: Exemplos de mamografias, onde é possível verificar o aparecimento de achados tumorais: (a) mamografia com o nódulo benigno selecionado e (b) mamografia com nódulo maligno selecionado. Fonte: (HEATH *et al.*, 1998).

por: gestação, uso de terapia de reposição hormonal e ganho ou perda de peso (DUARTE, 2006).

O tecido mamário é constituído principalmente por tecido granular e tecido de suporte ou conjuntivo (VOMWEG, 2008). O tecido granular é a parte responsável pela produção de leite durante o período de amamentação, sendo principalmente constituído de lóbulos e dutos. O tecido de suporte é principalmente constituído por tecido adiposo e por conectivos fibrosos que têm a função de manter a forma e sustentação da mama.

As glândulas lactíferas que compõe o tecido granular são circundadas por tecido adiposo e tecido conjuntivo. O leite secretado por elas flui através de canais até atingir o mamilo. Ao redor do mamilo, existe uma área externa de pele pigmentada denominada aureolo. As glândulas mamárias possuem um relacionamento com órgãos vizinhos, principalmente com o músculo grande peitoral, sobre o qual a mama descansa.

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 mamário vai se atrofiando e sendo

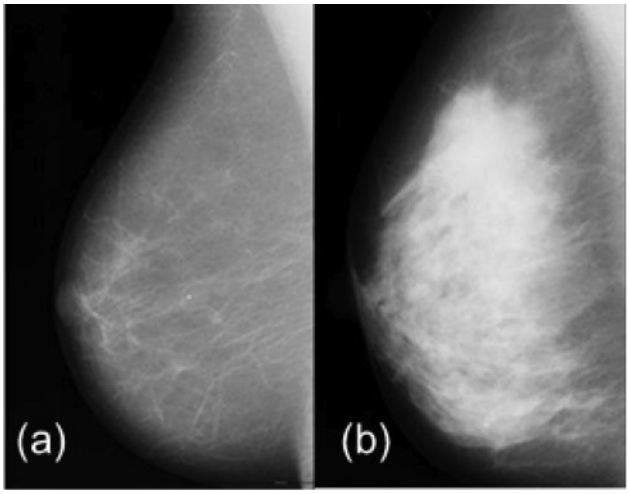


Figura 2.2: Exemplos de mamografias com densidade estruturais diferentes: (a) mama com baixa densidade e (b) mama com alta densidade.

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 (mama densa) e da mulher na pós-menopausa (mama lipo substituída com baixa densidade), como exemplificado pela Figura 2.2.

A densidade da mama é um dos aspectos preponderantes para a correta visualização da mesma. Em mamas muito densas, estruturas são dificilmente detectadas em imagens de mamografias devido a oclusão ocasionada pela sobreposição dos tecidos. Normalmente, pacientes com mamas mais densas são também jovens, para as quais é recomendado a utilização de exame de ultrassonografia que não é influenciada pela densidade da mama.

Qualquer uma das áreas da mama está sujeita a um distúrbio passível de ser visualizado através de imagens radiológicas (VOMWEG, 2008). Existem dois tipos principais de distúrbios da mama: achados benignos e malignos. O segundo é denominado de câncer por possuir condições de malignidade. As alterações benignas mais comuns estão associadas a tumores benignos, inflamações[1], cistos[2]

[1]As inflamações ou mastalgia são dores ou sensibilidade ao toque durante ou imediatamente antes do ciclo menstrual, provavelmente devido a alterações hormonais

[2]Os cistos são sacos cheios de líquido e que podem ser facilmente palpados. A causa dos

e alterações fibrocísticas. O câncer de mama é uma alteração maligna que pode ser classificado através de análise histopatológica em invasivo (ductais, tubulares, lobulares, medulares) ou não invasivo (ductal, carcinoma *in situ* e outros) (DUARTE, 2006).

Quando os achados correspondem a um volume de células resultante do crescimento desordenado, sem controle, são denominados também de neoplasias. Podem ser benignas ou malignas. O termo se deve à composição sólida que fica evidenciada na imagem pela não transparência em relação a incidência de raio-x.

Algumas neoplasias crescem muito lentamente e disseminam-se em outras partes do corpo somente após tornarem-se muito grandes. Outras são mais agressivas, crescendo e disseminando-se rapidamente. No entanto, o mesmo tipo de câncer pode evoluir de maneira diferente em mulheres diferentes. Apenas o médico que realizou a anamnese[3] e examinou a paciente pode analisar os aspectos específicos do câncer de mama apresentado pela mesma.

A maioria das mulheres apresenta, em algum momento de sua vida, nódulos de mama, geralmente localizados na área superior lateral, que são caracterizados como benignos e não representam maiores riscos de desenvolver o distúrbio de câncer. Desenvolver nódulos benignos é tão comum quanto apresentar dores na mama ou cistos. Logo, esta condição não se caracteriza como uma doença. Um tipo de nódulo mamário benigno são os fibroadenomas (nódulos fibrosos). Caracterizam-se por serem pequenos e sólidos recobertos por tecido fibroso e glandular. Estes nódulos geralmente ocorrem em mulheres jovens. São facilmente mobilizados, possuem bordas nitidamente definidas que podem ser apalpadas durante o auto-exame. Possuem consistência de borracha porque contêm colágeno.

O Carcinoma Ductal *in situ*, considerado raro no passado, passou a apresentar maior ocorrência depois da utilização de mamografias de rastreamento. O mesmo corresponde de 22 a 45% de todos os carcinomas de mama. O achado mamográfico mais comum que caracteriza este carcinoma é a presença de microcalcificações agrupadas com orientação ductal (DUARTE, 2006).

cistos mamários é desconhecida, embora eles possam ter relação com lesões

[3]Histórico colhida pelo médico sobre problemas de saúde do paciente. Importante para identificar grupos de risco e fazer o acompanhamento necessário. Dentre informações colhidas na anamnese estão: histórico de nódulos, retrações, desvios do mamilo, tempo de um achado clínico e velocidade de crescimento

As calcificações são pequenos depósitos de cálcio. São classificadas em Microcalcificações e Macrocalcificações. Macrocalcificações são depósitos maiores de cálcio causados por traumas, inflamações ou alterações regressivas de neoplasias benignas. Estas são encontradas na metade das mulheres acima de 50 anos e em 10% das mulheres abaixo dessa idade.

Microcalcificações são pequenos depósitos de cálcio que aparecem sozinhos ou em agrupamentos. Distribuição irregular em um ou mais agrupamento, além de contorno irregular, podem estar relacionado a características de malignidade (VOMWEG, 2008).

O Carcinoma Ductal Invasor representa em média 65% dos casos histologicamente comprovados e a maioria absoluta são nódulos palpáveis. Este carcinoma representa de 10 a 15% dos tumores malignos da mama e sua detecção é normalmente difícil. Seu aspecto é variado, desde espiculado quanto com presença de distorção arquitetural focal (DUARTE, 2006).

O prognóstico dos cânceres invasivos ductais e lobulares é similar. Outros tipos de neoplasias malignas menos comuns como, por exemplo, o carcinoma medular e o carcinoma tubular[4], apresentam um prognóstico um pouco melhor. Normalmente é através de imagens de radiografia que os achados clínicos são encontrados e acompanhados. Essas imagens são obtidas através de programas de rastreamento que têm como objetivo principal encontrar o distúrbio em fase inicial, aumentando as chances de cura e o tempo de sobrevida.

Devido a grande variedade de comportamento que uma neoplasia assume, o Colégio Americano de Radiologia, junto com outros órgãos, criou o BI-RADS (ACR, 2003b), uma classificação para achados na mama quanto ao risco de malignidade e também padronizou algumas classificações para o tipo de composição do tecido da mama. O objetivo do BI-RADS criado na década de 90 foi de uniformizar o laudo médico, padronizar os termos empregados, estabelecer categorias de avaliação final e sugerir condutas apropriadas para cada uma delas.

O BI-RADS classifica a composição mamária quanto as quantidades relativas de tecido adiposo e granular. São quatro classificações utilizadas como padrão:

1. Mamas predominantemente adiposas (25% do componente granular);

[4]Originados nas glândulas lactíferas

2. Mamas parcialmente adiposas (com densidades de tecido granular ocupando de 26% a 50% do volume da mama);

3. Mamas com padrão denso e heterogêneo (51% a 75% de tecido granular, dificultando a percepção de nódulos) e

4. Mamas muito densas, apresentando mais de 75% de tecido granular (diminuindo a sensibilidade da mamografia).

Já os achados radiográficos, exemplificados pela Figura 2.3, são descritos como:

- Massas: qualquer opacidade com algum contorno arredondado e definido segundo a forma e a densidade;

- Microcalcificações agrupadas: conjunto de pequenos depósitos de cálcio, classificados de acordo com sua morfologia e distribuição e

- Distorção focal de arquitetura: espiculações em uma região da mama ou uma retração focal do contorno parenquimatoso denso.

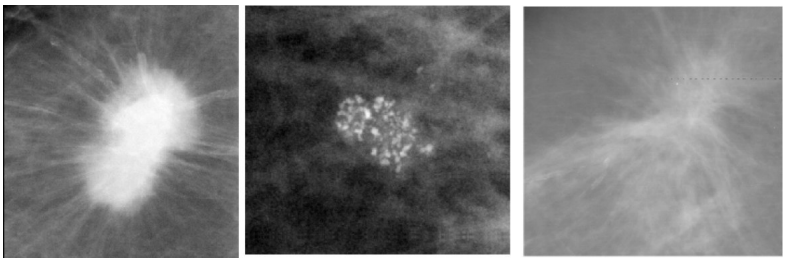


Figura 2.3: Anormalidades encontradas em tecidos mamários. Da esquerda para direita: massa espiculada, agrupamento de microcalcificações, distorção de arquitetura. Fonte: (HEATH *et al.*, 1998).

O BI-RADS também classifica os graus de malignidade do achado clínico em seis categorias conforme Tabela 2.1[5].

[5]*Menor, médio e maior respectivamente, **Valor Preditivo Positivo

Tabela 2.1: Categorização BI-RADS dos achados suspeitos.

Categoria	Interpretação	VPP**	Conduta
0	Inconclusivo		Exame adicional
1	Exame normal	0%	Controle anual a partir dos 40 anos
2	Achados benignos, como calcificações, linfonodos intramamários, cistos, etc;	0%	Controle anual a partir dos 40 anos
3	Provavelmente benigno	< 2%	Controle semestral
4 (A,B,C)*	Suspeito	>2% e < 90%	Biópsia
5	Provavelmente Maligno	>95%	Biópsia
6	Lesão maligna – biopsada ou diagnosticada não submetida a terapia intensiva	100%	

2.1.1 Imagens da Mama

Três tipos de imagens da mama são comumente utilizadas: mamografia, ultrassonografia e ressonância magnética.

A mamografia é atualmente o único exame com potencialidade comprovada para detectar o câncer clinicamente oculto da mama em tamanho e estadiamento precoces. Por isso, também é o único exame capaz de reduzir a mortalidade através do rastreamento de mulheres assintomáticas (KOPANS, 2007).

Grande parte da estrutura da mama é formada por tecido adiposo, que é radiolucente. Isto significa que esse tipo de tecido é permeável à incidência de raio-x. Por outro lado, os principais componentes da densidade radiográfica na mamografia são os tecidos conjuntivos, que são responsáveis pela maior parte das variações grosseiras de densidade. A Figura 2.4 apresenta exemplo de imagem de mamografia após sua digitalização.

O objetivo da mamografia é produzir imagens de alta resolução das estruturas

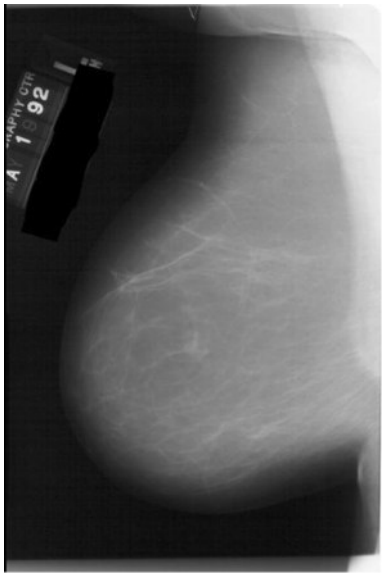


Figura 2.4: Exemplo de mamografia, na posição MLO, obtida da base de dados *Digital Database for Screening Mammography* - DDSM. Fonte: (HEATH *et al.*, 1998).

internas da mama, a fim de permitir a detecção do câncer de mama. Os dois tipos principais de lesões que podem ser visualizadas através de uma mamografia são calcificações e massas. 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.

Por isso, a mamografia deve ser realizada em um aparelho de raio-x específico, chamado mamógrafo. Nele, a mama é comprimida de forma a fornecer melhores imagens e, portanto, melhor capacidade de diagnóstico. A compressão é necessária para evitar a subexposição da base e a superexposição dos tecidos anteriores da mama, mais finos.

A mamografia é bilateral, ou seja, é feita uma radiografia de cada mama. Além disso, em um exame de mamografia, duas incidências ou projeções de cada mama são indispensáveis: uma visão médio-lateral oblíqua (MLO) e uma crânio-caudal (CC) (Figura 2.5).

A incidência médio-lateral oblíqua é a mais útil, pois permite a visualização do alto da axila para baixo, incluindo a prega infla-mamária. O termo oblíquo não

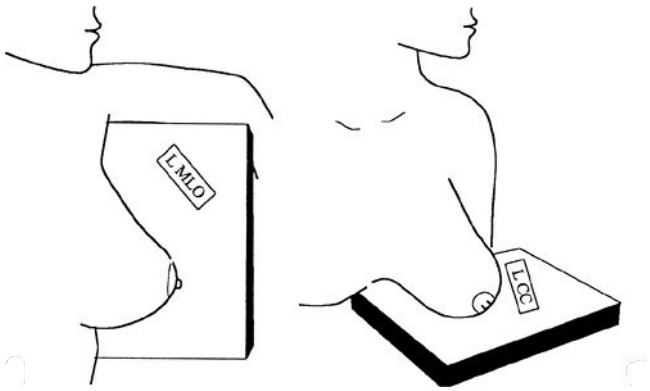


Figura 2.5: Principais incidências de uma exame de mamografia. À esquerda, incidência MLO e à direita, incidência CC. Fonte: adaptado de (KOPANS, 2007).

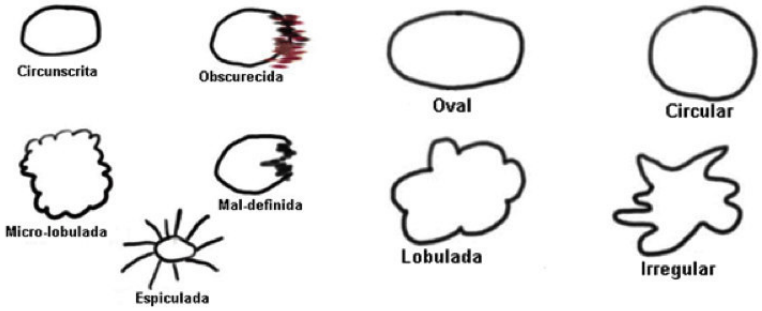
se aplica à paciente, mas ao plano de compressão da mama. O músculo peitoral deve ser visível, estendendo-se obliquamente até a metade superior da imagem. Além disso, deve ser muito largo no alto e ir se afilando à medida que cruza a parte superior da mama. Estudos sugerem que a mama é representada de maneira ótima quando o músculo peitoral é visível até o eixo do mamilo.

A projeção crânio-caudal 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 assim a visão médio-lateral oblíqua.

A imagem produzida através da mamografia convencional é analógica, ou seja, os componentes de uma imagem em filme são contínuos. Duas abordagens são utilizadas na obtenção de imagens de mamografias digitais: a digitalização de mamografias convencionais de filme/película e a aquisição direta (ou primária) de dados digitais da mama por detectores em estado sólido.

Dentre as várias anormalidades que podem ser detectadas através da mamografia, as massas, que correspondem ao objeto de estudo deste trabalho, representam o tumor em si e aparecem como regiões densas, de tamanho e formato variáveis. Podem ser classificadas de acordo com o aspecto de suas bordas, como circunscritas, microlobuladas, obscurecidas, mal definidas e espiculadas (Figura 2.6a). Com relação ao formato, podem ser classificadas em redondas, ovais, lobulares ou irregulares. (Figura 2.6b).

As massas benignas são geralmente bem definidas e com pouca repercussão na



(a) (b)

Figura 2.6: Classificação das massas: (a) baseado no aspecto de suas bordas. (b) baseado na sua forma. Extraído de Mitchell *et al.* (2008).

arquitetura vizinha da mama. Massas que possuem fronteiras irregulares, e que se confundem com tecidos adjacentes, têm grande probabilidade de terem caráter maligno (VALE, 2002). Entretanto, o diagnóstico de massas através da mamografia não é tão simples na maioria dos casos, a menos que as massas apresentem sinais característicos de um processo maligno ou, no caso das massas benignas, calcificações benignas e/ou bordas encapsuladas. Por essa razão, a mamografia é considerada uma excelente técnica de detecção (rastreamento), porém não é tão boa para a realização de diagnósticos (definição de caráter benigno ou maligno de uma massa). Cabe também ressaltar que o tumor de mama nem sempre produz uma massa mamograficamente visível. Ele pode produzir somente uma área de distorção de arquitetura.

Outra característica comum é que o fluxo das estruturas na mama é dirigido para o mamilo. Os distúrbios nesse fluxo devem ser avaliados com cuidado, embora possam ser causados por fatores benignos (KOPANS, 2007).

A ultrassonografia da mama normalmente é realizada para avaliar achados ambíguos da mamografia ou exame físico. É uma tecnologia mais barata e mais acessível que a mamografia ou ressonância magnética. Geralmente é utilizada em uma área bem específica da mama, previamente selecionada, conforme exemplificado pela Figura 2.7. É indicada para detecção de massas sendo a maneira mais simples de determinar se uma massa não passa de um cisto sem ter que realizar um procedimento de biopsia. Ainda é indicada para determinar se um

tumor é benigno ou maligno, especialmente em mulheres jovens que normalmente possuem mamas mais densas. Essa última característica faz com que a imagem de ultrassom seja utilizada como primeiro passo para o diagnóstico de achados clínicos na mama (VOMWEG, 2008).

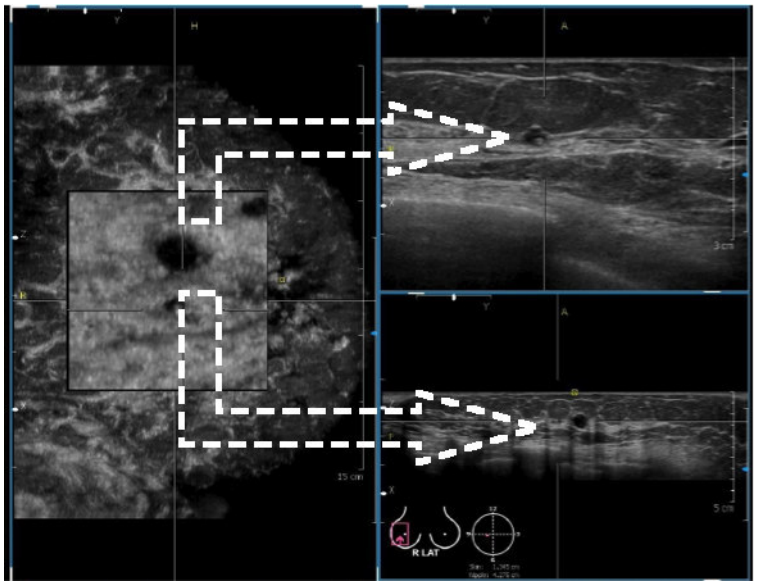


Figura 2.7: Um exemplo de imagem de ultrassom e sua relação com o exame de mamografia. À esquerda uma mamografia com um achado clínico a ser investigado. À direita duas imagens de ultrassom da região selecionada. O achado aparece como uma região circular e escura. Fonte: adaptado de (USYSTEMS, 2013).

A ressonância magnética da mama apresenta a imagem com mais alto grau de qualidade e contraste entre as imagens radiológicas da mama normalmente realizadas em programas de rastreamento. A imagem é na verdade um volume composto por fatias ao longo do tempo formando uma imagem 3D. É o método fundamental para pacientes que possuam implantes de silicone, pacientes em processo de quimioterapia e do grupo de alto risco, além de ser utilizada para planejamento cirúrgico (VOMWEG, 2008).

O processo de obtenção da imagem é realizado em duas etapas. Na primeira a região analisada é varrida e são criadas as fatias. Na segunda, é realizada uma nova varredura após a injeção do agente de contraste paramagnético usando os

mesmos parâmetros da primeira captura. Por fim, as duas imagens são subtraídas para gerar uma nova imagem, realçando estruturas com alta homogeneidade. O resultado final pode ser obtido por fatia e é possível analisar com boa precisão aspectos morfológicos da região de interesse. Um exemplo de uma fatia segmentada é apresentada na Figura 2.8.

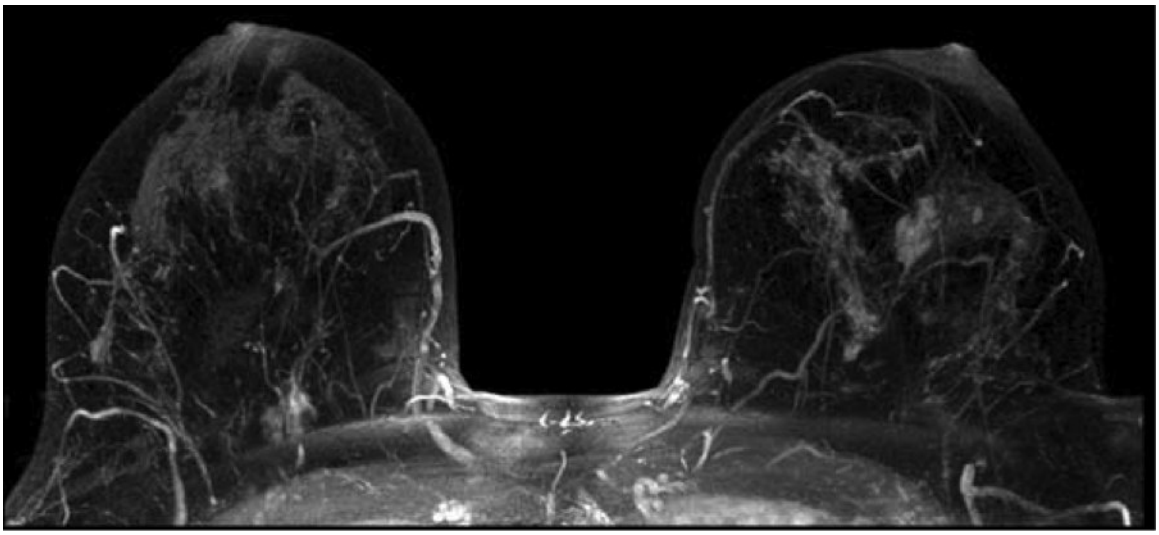


Figura 2.8: Exemplo de fatia de imagem de ressonância magnética. Extraído de RSNA (2008).

Nas próximas seções desse capítulo são apresentados os conceitos teóricos utilizados pela metodologia proposta para a detecção automatizada de massas em imagens de mamografias. Os aspectos fisiológicos das massas expostos nessa seção serão levados em consideração para o aperfeiçoamento das técnicas propostas na metodologia apresentada no próximo capítulo.

2.2 Processamento de Imagens Digitais

O processamento de imagens digitais é definido como sendo o conjunto de técnicas computacionais que transformam uma imagem digital de entrada em uma saída desejada, normalmente outra imagem. Dessa maneira, é possível melhorar o aspecto visual de certas feições estruturais para o observador humano e fornecer outros elementos para a interpretação visual da imagem, podendo

inclusive gerar outros produtos que possam ser posteriormente submetidos a outros processamentos.

Historicamente, o interesse em métodos de processamento imagens digitais surgiu, principalmente, da necessidade de melhorar a qualidade da informação presente na imagem para interpretação humana. A evolução da tecnologia de computação digital, bem como o desenvolvimento de novos algoritmos para lidar com sinais bidimensionais, permitem uma atuação maior do processamento de imagens em áreas transversais à ciência da computação, como por exemplo: medicina, astronomia, arqueologia, arquivologia, biologia, dentre outras.

Embora o principal objetivo do processamento de imagens seja o auxílio à compreensão da mesma, existem uma vasta gama de algoritmos com finalidade muito específicas, que juntos compõe a metodologia final. Uma metodologia difere de outra na maneira que compõe seus passos ou ferramentas, mas tipicamente obedecem as etapas apresentadas por Gonzalez e Woods (2010), conforme Figura 2.9.

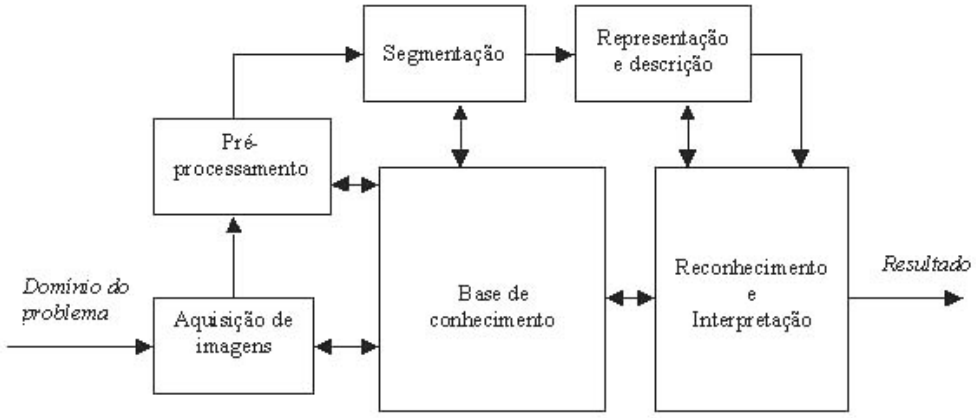


Figura 2.9: Passos fundamentais em processamento de imagens digitais. Fonte: adaptado de (GONZALEZ; WOODS, 2010).

Dentre as etapas necessárias após a definição e delimitação do problema, estão: aquisição das imagens digitais; pré-processamento; segmentação; representação; descrição e, reconhecimento e interpretação. O conjunto de resultados gerados por uma etapa é utilizado na etapa seguinte. Algumas etapas não recebem ou geram imagens digitais como resultados. Assim, ao final de todo o processamento,

o resultado pode ou não ser caracterizado por uma imagem digital. Também cabe lembrar que uma metodologia pode conter apenas um subconjunto de todas as etapas apresentadas.

No restante deste capítulo são abordados os aspectos teóricos das técnicas utilizadas na proposta de metodologia desta tese. Para melhor compreensão, cada um destes métodos está organizado em etapas conforme a divisão das etapas fundamentais do processamento de imagens.

2.3 Pré-Processamento

O realce da imagem é um processo importante para o bom desempenho dos algoritmos de segmentação, extração de características e reconhecimento de padrão. Para o desenvolvimento da metodologia proposta por este trabalho, foram utilizadas técnicas de realce para o tratamento do contraste de imagens digitalizadas de mamografia. As técnicas utilizadas são: Suavização por Média (GONZALEZ; WOODS, 2010), Realce Logarítmico, *Constrast Streching* e *Contrast-Limited Adaptive Histogram Equalization* (CLAHE) (PIZER, 1987). Devido ao uso generalizado, a primeira técnica não é descritas nas seções seguintes.

2.3.1 *Contrast-Limited Adaptive Histogram Equalization* (CLAHE)

Técnicas de contraste adaptativo produzem bons resultados no realce de imagens médicas. Todavia, ao mesmo tempo que estruturas de interesse são realçadas, sinais de ruído também são realçados (PIZER, 1987). Os ruídos em excesso podem influenciar na interpretação dos resultados. É adequado que as técnicas de contraste adaptativo possuam um mecanismo de limitação de contraste para regular o resultado final.

O *Contrast-Limited Adaptive Histogram Equalization* (CLAHE) (PIZER, 1987) é um algoritmo que divide a imagem em regiões contextuais e aplica a equalização de histograma a cada região individualmente. Isso equilibra a distribuição de valores de cinza utilizados e, assim, torna as características ocultas da imagem mais visíveis.

O algoritmo pode ser implementado com diferentes formas de funcionamento. A seleção da região pode ser feita utilizando bordas sobrepostas ou não. A região pode ser baseada na detecção de bordas. A mesma região pode ainda ter tamanho variado, dependendo da aplicação e o histograma pode estar associado a uma quantia máxima de faixas.

A forma padrão do CLAHE utiliza um método uniforme de realce:

$$g = (x_{max} - x_{min})H(x) + x_{min} \tag{2.1}$$

onde g representa o novo valor do pixel, x o valor original do pixel, x_{max} e x_{min} os valores de pixel máximo e mínimo, respectivamente, informados baseado na faixa máxima de valores da imagem. Por último, a função H representa o histograma da região em análise. Verifica-se que pela expressão original do CLAHE, apenas uma equalização de histograma local é realizada.

Variações da implementação original propõem a limitação do contraste local segundo um parâmetro α nas expressões. Esse parâmetro é um ponderador de contraste máximo a ser atribuído. Uma das implementações existentes que regulam dessa maneira o contraste é a exponencial:

$$g = x_{min} - \frac{1}{\alpha}\ln(1 - H(x)) \tag{2.2}$$

onde α representa uma constante de adaptação de contraste. A Figura 2.10 apresenta um exemplo da utilização do CLAHE no processamento de imagens de mamografia.

2.3.2 Realce Logarítmico

A transformação logarítmica de uma imagem mapeia uma faixa estreita de baixos valores de intensidade de entrada em uma faixa mais ampla de níveis de saída (GONZALEZ; WOODS, 2010). O oposto se aplica aos valores mais altos de níveis de intensidade de entrada. Utilizamos uma transformação desse tipo para expandir os valores de pixels mais escuros em uma imagem ao mesmo tempo em que comprimimos os valores de nível mais alto.

A transformação é realizada através da seguinte equação:

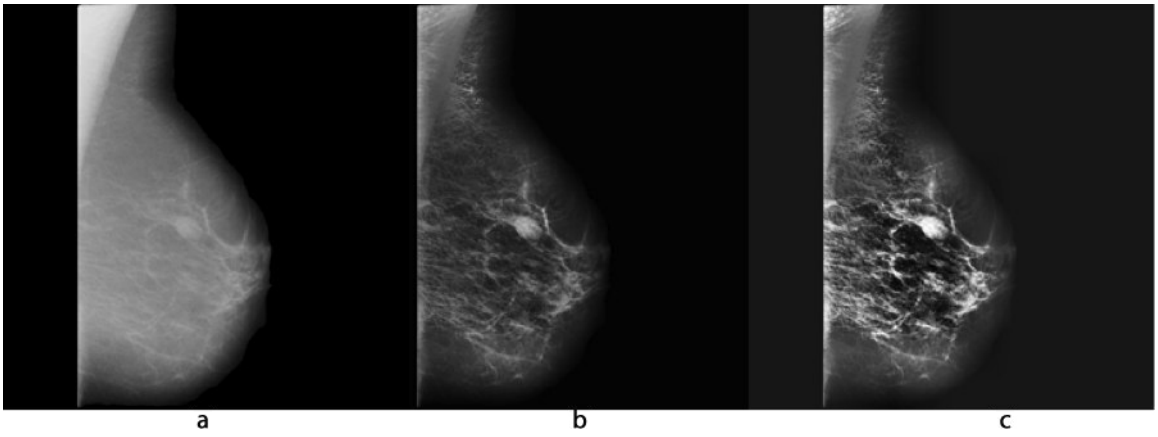


Figura 2.10: Exemplo de realce de imagem usando CLAHE: (a) imagem original, (b) imagem processada por CLAHE usando $\alpha = 0.018$, e (c) imagem processada pelo CLAHE usando $\alpha = 0.3$.

$$s = c \log(1 + r) \tag{2.3}$$

onde c é uma constante arbitrada como o tamanho máximo do canal, s é o novo valor calculado e r o valor original da imagem. A Figura 2.11 mostra uma imagem de uma massa maligna antes e após o realce logarítmico com $c = 255$.

2.3.3 *Contrast Stretching*

Esta técnica consiste em melhorar o contraste da imagem de maneira dinâmica, através da seleção de parâmetros de adaptação. O objetivo principal é aumentar a faixa dinâmica de tons de cinza da imagem e assim realçar o contraste (GONZALEZ; WOODS, 2010). Portanto, essa função pode ser aplicada após operações que agrupem tons de cinza, reduzindo o contraste da imagem.

O método se resume à definição de dois pontos r_1 e r_2 para controlar a forma da transformação. Após a escolha do limiares, basta seguir a relação: valores entre r_1 e r_2 de uma distribuição de tons linear e valores menores que r_1 e maiores que r_2 sofrem limiarização. No caso da escolha de parâmetros de corte r_1 e r_2 equivalentes ao menor e maior tom de cinza, respectivamente, o método provoca a redistribuição dos tons de cinza ao longo da faixa completa de valores.

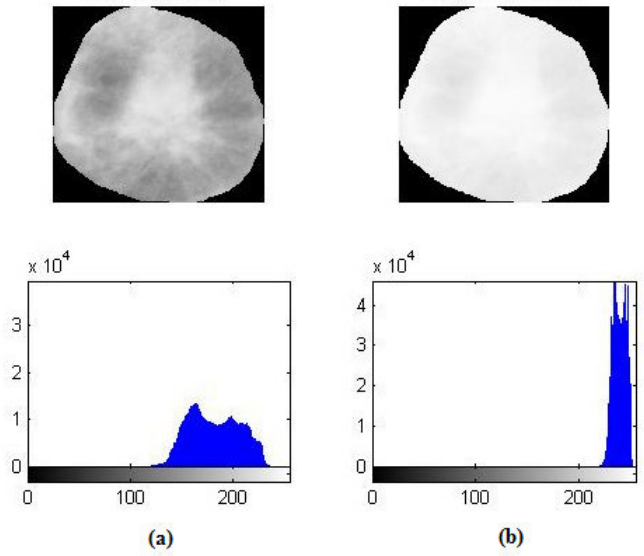


Figura 2.11: Exemplo da aplicação do realce logarítmico em uma massa maligna: à esquerda antes do realce logarítmico e à direita após o realce logarítmico. O histograma abaixo de cada imagem demonstra como as tonalidades foram reajustadas para tonas mais altos.

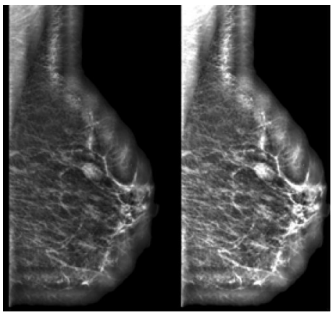


Figura 2.12: Exemplo de utilização do *Contrast Stretching*. À esquerda verificamos uma mamografia com baixo contraste embora as estruturas estejam separáveis. À direita, após o método, as estruturas são realçadas e os tons de cinza distribuídos sobre toda a faixa de valores

2.4 Segmentação

A segmentação consiste na identificação de objetos numa imagem. Este é um processo dependente da classe de imagens que se está trabalhando. As seções

seguintes apresentam os algoritmos de *MeanShift* para agrupamento inicial de objetos semelhantes na imagem baseado em densidade e o *Fast Scanning* para separação e identificação de objetos baseado no crescimento de regiões, ambos utilizados pela metodologia para detecção de regiões suspeitas.

2.4.1 *MeanShift*

MeanShift (CHENG, 1995) é um algoritmo não paramétrico para localização de máximos locais a partir de dados discretos amostrados de uma determinada função ou distribuição de pontos. Quando usado como agrupador, sua funcionalidade pode ser definida como uma função que estima o conjunto de máximos de um grupo de pontos e em seguida faz com que cada ponto assuma um dos máximos como seu representante, provocando o agrupamento dos dados. Uma de suas vantagens é que não é necessário ter anteriormente informações de forma ou quantidade de objetos que se deseja encontrar.

Logo, dada uma coleção de n pontos $x_i, i = 1, ..., n$ num espaço d-dimensional R^d, o *kernel* $K(x)$ é utilizado para estimar a função de densidade dos pontos sobre uma janela de raio h, obtida pela equação:

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{2.4}$$

onde d representa a dimensão do espaço de características dos pontos utilizados. A função *kernel* utilizada pode variar de acordo com a necessidade imposta pelo problema. Neste trabalho utilizamos a função gaussiana definida pelo *kernel*:

$$K(x) = e^{-\frac{x^2}{2\sigma^2}} \tag{2.5}$$

Para cada ponto é realizada a ascendência do gradiente sobre a densidade local estimada até que se encontre a maior densidade local. Os pontos estacionários encontrados pela ascendência representam modos de densidade. Todos os pontos associados a um mesmo modo é chamado de agrupamento. Então, assumindo que $g(x) = -K(x)$, a Equação 2.6 define a função $m(x)$ chamada de *MeanShift*:

$$m(x) = \frac{\sum_{i=1}^{n} g(\frac{x-x_i}{h})x_i}{\sum_{i=1}^{n} g(\frac{x-x_i}{h})} - x \tag{2.6}$$

O funcionamento do algoritmo consiste em definir para cada ponto um círculo envolvente de raio h ao redor do ponto e calcular a média do mesmo, dado pelo vetor $m(x_i^t)$. Em seguida, desloca-se o centro do círculo pela variação de média calculado por $m(x_i^t)$. O algoritmo repete esses passos até convergir. Após cada iteração, podemos considerar que a janela se desloca para uma região mais densa do conjunto de dados.

2.4.2 *Fast Scanning Algorithm* (FSA)

O *Fast Scanning Algorithm (FSA)* é uma implementação alternativa de algoritmo de crescimento de regiões que não necessita de sementes como ponto de partida. O conceito do algoritmo é analisar todos os pontos da imagem a partir do canto superior esquerdo em direção ao canto inferior direito e determinar quando é necessário realizar a junção de um novo ponto com os pontos vizinhos existentes ou criar um novo grupo. O critério de junção é dado por um corte ou distância calculada do novo elemento em relação ao grupo existente.

Basicamente, o algoritmo segue os passos:

1. Com o primeiro pixel P_i da imagem, na posição (1,1), criar o primeiro grupo C_1

2. Continuar analisando os próximos pixels (P_j) e determinar se cada ponto deve ser unido com o grupo mais a esquerda e superior, ou criar um novo grupo. A decisão depende da média do grupo e da distância deste em relação ao novo elemento, dado pelo limiar de corte, conforme abaixo:

    - Se (P_j - media(C_i)) $\leq$ limiar então: unir o ponto com o grupo e recalcule a média
    - Se (P_j - media(C_i)) $>$ limiar então: criar um novo grupo C_j com o ponto P_j

- Obs.: se houver mais de um grupo onde a distância em relação a média é menor que o corte, se escolhe o grupo que tiver a menor distância.

3. Repetir os passos anteriores até que todas os pontos tenham sido agrupados

4. Remover todos os grupos pequenos: verificar se existem grupos C_i com uma quantidade menor que X pontos. Caso exista, procurar um grupo vizinho a C_i com a menor distância entre as médias. Unir o grupo pequeno com o vizinho mais similar.

De uma maneira geral, o FSA é uma rotina simples de agrupamento baseado em crescimento de regiões, escolhido principalmente pela facilidade de se promover extensões em seu algoritmo.

2.5 Extração de Características

A descrição de imagens é realizada como um passo anterior ao reconhecimento de padrões. O objetivo desta etapa é quantificar e mensurar um padrão perceptível ou não visualmente. O padrão é abstraído como uma assinatura e tem como finalidade a categorização de objetos nas imagens. A descrição é realizada sobre aspectos de textura ou morfologia do objeto de interesse.

A textura é uma propriedade importante na percepção de regiões e superfícies, contendo informações sobre a distribuição espacial das variações de tonalidade locais em valores de pixels que se repetem de maneira regular ou aleatória ao longo do objeto ou imagem. Comumente é definida em termos de uniformidade, densidade, aspereza e intensidade, dentre outras. A textura é caracterizada como um conceito bidimensional, onde uma dimensão contém as propriedades primitivas da tonalidade e a outra corresponde aos relacionamentos espaciais entre elas.

Dentre algumas das abordagens de análise de textura estão as abordagens baseadas em análise de diversidade das tonalidades e estatística espacial. A primeira quantifica textura em termos de relacionamento de espécies que representam tonalidades de pixels da imagem. A segunda quantifica os relacionamentos espaciais entre os pixels e suas tonalidades.

Além da textura, os objetos em uma imagem também podem ser caracterizados conforme seus aspectos de forma. A análise das formas dos objetos é bastante

utilizada quando é possível quantificar os diferentes aspectos de formato de cada objeto. A mesma pode ser feita através de geometria convexa a qual leva em consideração os aspectos de forma externa do objeto. E ainda com geometria côncava que além de quantificar as formas externas também realiza a descrição de geometria interna.

2.5.1 Correlação de Histograma

O histograma é a base para muitas técnicas de pré-processamento e descrição de textura no domínio espacial da imagem. O histograma de uma imagem digital com intensidades variando de 0 a $L-1$ é obtido pela função discreta $f(r_k) = n_k$ onde r_k é a k-ésima intensidade e n_k é o número de *pixels* da imagem com a intensidade r_k (GONZALEZ; WOODS, 2010).

Apenas a informação de histograma isolado pode trazer distorções quanto à ocorrência de tonalidades quando as imagens analisadas não possuem a mesma dimensão espacial. Para evitar tais distorções, são realizadas operações de normalização de histograma. A normalização é realizada com a divisão da frequência acumulada pelo maior valor de frequência para todas as tonalidades.

A partir do histograma normalizado se torna possível o cálculo de correlação entre dois histogramas de duas regiões distintas para identificar o grau de associação das mesmas, utilizando a seguinte equação de correlação:

$$d(F_1, F_2) = \frac{\sum_I (F_1(I) - \bar{F_1})(F_2(I) - \bar{F_2})}{\sqrt{\sum_I (F_1(I) - \bar{F_1})^2 (F_2(I) - \bar{F_2})^2}} \tag{2.7}$$

onde F representa o histograma sobre as tonalidades I. O valor informado pela correlação varia de -1 a 1, sendo que quanto mais próximo a 1 maior a similaridade das texturas. A correlação de histograma é utilizada nessa proposta como um filtro de regiões que tenham pouca correlação com regiões de alta intensidade.

2.5.2 Análise de Diversidade

A Análise de Diversidade, utilizada na Ecologia para medir a biodiversidade de um ecossistema, pretende identificar a distribuição de um grupo de espécies e as suas relações. A diversidade se refere à variedade de espécies em uma dada

comunidade ou *habitat*. A biodiversidade é a relação entre o número de espécies (riqueza), o padrão de distribuição dos indivíduos nas suas espécies (uniformidade) e o domínio de uma ou mais espécies sobre as outras (dominância). Todas essas características podem ser medidas e investigadas através do uso de índices geralmente classificados quanto a cobertura local (alfa) ou entre vários *habitats* (beta) (MAGURRAN, 2004).

De modo mais geral, os índices de diversidade podem ser usados para medir a diversidade de uma população em que cada membro pertence a um único grupo ou espécie. Adotamos que os pixels são os indivíduos e suas tonalidades representam o conjunto de espécies presentes no ecossistema.

Considerando-se que cada região tem uma distribuição de tons de cinza que varia de 0 a 255 (8 bits por pixel). Assim, qualquer pixel x da imagem A tem uma espécie S_i equivalente a sua tonalidade. O conjunto dado por $x_0, x_1, ..., x_N$ representa a população total P, onde N é o número total de indivíduos e também de pixels. S representa a quantidade total de espécies do conjunto $s_0, s_1, ..., s_i$; onde i representa uma espécie específica. O número de indivíduos de cada espécie é representada por n_i; p_i é a proporção total da amostra pertencente à espécie i, calculado por $p_i = \frac{n_i}{N}$.

O índice de Shannon-Wienner (SHANNON, 2001) é derivado da Teoria da Informação, mostrando o grau de incerteza que existe em relação às espécies de um indivíduo escolhido aleatoriamente de uma população. O cálculo é definido por:

$$H = -\sum_{i=1}^{S} p_i \ln p_i \tag{2.8}$$

Um característica do índice de Shannon-Wienner é que não é necessário conhecer anteriormente a distribuição da população inteira de espécies para usá-lo. Espécies raras e abundantes têm pesos iguais. Os maiores valores do índice representam maior heterogeneidade da população em estudo e consequentemente maior riqueza. Além disso, valores semelhantes, tomadas a partir de populações separadas, representam uniformidade sobre todas as espécies.

O índice de McIntosh (MAGURRAN, 2004) trata uma população como um ponto S-dimensional de um hipervolume[6]. A distância euclidiana da comunidade para

[6]Volume multidimensional onde cada indivíduo de cada espécie é projetado tendo como

a origem é usada como uma medida da diversidade, definida por:

$$Mc = \frac{N - U}{N - \sqrt{N}} \tag{2.9}$$

onde $U = \sqrt{\sum_{i=1}^{S} {n_i}^2}$ é uma distância euclidiana desde a origem até a população. O valor do índice varia de 0, se houver baixa diversidade, a 1 quando a diversidade é máxima.

O índice de Diversidade Total (MAGURRAN, 2004) estima a riqueza total de uma população com base na variação de espécies. Esta medida é obtida por:

$$Td = \sum_{i=1}^{S} w_i(p_i(1 - p_i)) \tag{2.10}$$

onde w_i representa a importância proporcional de cada espécie sendo calculado por $\frac{1}{n_i}$.

O índice de Brillouin (PIELOU, 1975) mede a riqueza de uma população conhecida, sendo recomendado quando a população não é aleatória. Além disso, este índice tende a informar resultados semelhantes ao índice de Shannon-Wienner quanto este é usado em uma população que não é completamente conhecida. Este índice é definido por:

$$Eb = \left(\frac{1}{N}\right)(\log N! - \sum_{i=1}^{S} \log n_i!) \tag{2.11}$$

O índice de Simpson é uma medida de segunda ordem estatística que informa a probabilidade de dois indivíduos escolhidos aleatoriamente de uma comunidade pertencerem à mesma espécie (SIMPSON, 1949). Sua principal funcionalidade é resumir a representação desta diversidade em um único valor capaz de qualificar a região como muito heterogênea ou uniforme. Sua equação é obtida por:

$$Ds = \frac{\sum_{i=1}^{S} n_i(n_i - 1)}{N(N - 1)} \tag{2.12}$$

Assim como Shannon-Wienner, o índice de Simpson leva em conta a riqueza de espécies sendo uma medida da abundância relativa de cada espécie (HILL, 1973). Os valores obtidos para o índice de Simpson estão no intervalo de 0 a 1, onde

referência o ponto central ou origem.

o valor 0 representa a diversidade infinita na amostra e 1 significa que não há nenhuma diversidade.

O índice de Berger-Parker (MAY, 1975) mede a importância numérica da espécie mais abundante em relação a toda a população, definido por:

$$Bp = \frac{max(n_i)}{N} \tag{2.13}$$

onde $max(n_i)$ é a espécie mais abundante dentro da população. O objetivo do índice é representar a dominância da espécie mais abundante em relação a todo o ecossistema.

O índice J (PIELOU, 1975) utiliza o índice de Shannon-Wienner para obter a distribuição de indivíduos entre as espécies observadas que maximizam a diversidade. Definido por:

$$J = \frac{H}{H'} \tag{2.14}$$

onde H é o índice de Shannon-Wienner (Equação 2.8) e H' o seu máximo obtido por:

$$H' = \log S \tag{2.15}$$

O índice Ed (MAGURRAN, 2004) compara a dominância de Simpson com o conjunto de espécies conhecidas de maneira a maximizar a diversidade. É calculado através da relação do índice de Simpson e seu valor máximo:

$$Ed = \frac{Ds}{Ds'} \tag{2.16}$$

onde Ds' é calculado por:

$$Ds' = \left(\frac{S-1}{S}\right)\left(\frac{N}{N-1}\right) \tag{2.17}$$

O índice de Hill (JOST, 2010) calcula a uniformidade da distribuição de espécies, baseando-se na relação de dominância e diversidades presentes na análise. O índice é obtido por:

$$Hill = \frac{\frac{1}{Ds-1}}{e^H - 1} \tag{2.18}$$

onde Ds é o índice de Simpson (Equação 2.12) e H representa o índice de Shannon-Wienner (Equação 2.8).

O índice de Buzas-Gibson (BUZAS; HAYEK, 1998) indica o grau de uniformidade utilizando o índice de Shannon-Wienner, dado por:

$$Bg = \frac{e^H}{S} \tag{2.19}$$

O índice de Camargo (CAMARGO, 1993) é um índice de uniformidade que não considera riqueza das espécies, assim não é afetado por espécies raras na população. Toma como base a proporção relativa da espécie i em relação as demais espécies, sendo definido por:

$$E = 1 - \left(\sum_{i=1}^{S} \sum_{j=i+1}^{S} \frac{p_i - p_j}{S} \right) \tag{2.20}$$

2.5.3 Análise Espacial

A análise espacial é um estudo quantitativo de fenômenos localizados no espaço. Dessa forma, os índices utilizados em estatística espacial analisam a informação em termos de localização espacial, ou seja, o fenômeno estudado possui alguma forma de localização. Muitos dados de uso comum possuem alguma referência espacial como, por exemplo, dados censitários, sempre relacionados ao local de residência do indivíduo. Sob esta ótica, muitos dados que podem ser analisados estatisticamente possuem referência espacial.

A estatística espacial traz resultados diferentes daqueles obtidos pela estatística clássica. Para sua análise são necessárias, pelo menos, as informações sobre a localização e os atributos, que são valores associados aos dados independentemente da forma como sejam medidos. Parte-se do pressuposto que os dados são espacialmente dependentes.

Uma das taxonomias mais utilizada para caracterizar os problemas de análise espacial considera três tipos de dados (CÂMARA, 2003):

- Eventos ou Padrões Pontuais: fenômenos expressos por ocorrências identificadas como pontos localizados no espaço, denominados processos pontuais. São exemplos: localização de crimes, ocorrências de doenças e de espécies vegetais. O objeto de interesse é a própria localização espacial

dos eventos em estudo. O objetivo é estudar a distribuição espacial desses pontos (se é aleatório ou não), se contém aglomerados ou está regularmente distribuído, ou estabelecer o relacionamento de ocorrência de eventos com características individuais.

- Superfícies contínuas: estimada com base em um conjunto de amostras de campo que podem estar regularmente distribuídas. O objetivo é reconstruir a superfície da qual se retirou e mediu as amostras. Usualmente, esse tipo de dado é resultante do levantamento de recursos naturais e inclui mapas geológicos, topográficos, ecológicos, fitogeográficos e pedagógicos; e

- Áreas com Contagens e Taxas Agregadas: trata-se de dados associados a levantamentos populacionais, como censos e estatísticas de saúde e que originalmente referem-se a indivíduos localizados em pontos específicos do espaço. Esses dados são agregados em unidades de análise, usualmente delimitadas por polígonos fechados (setores censitários, zonas de endereçamento postal, municípios), onde se supõe existir homogeneidade interna.

O processo de análise de pontos pode ser descrito em termos dos efeitos de primeira e segunda ordem. Os efeitos de primeira ordem, considerados globais ou de grande escala, correspondem a variações no valor médio do processo no espaço. Efeitos de segunda ordem, denominados locais ou de pequena escala, representam a dependência espacial no processo proveniente da estrutura de correlação espacial. Análise de segunda ordem trata um número maior de vizinhos visualizados através dos vizinhos mais próximos.

O processo de análise de dados espaciais contém métodos de visualização, métodos exploratórios para investigar algum padrão nos dados e métodos que auxiliem a escolha de um modelo estatístico e a estimação dos parâmetros desse modelo. Dessa forma, as estatísticas de segunda ordem usadas para descrever tanto pontos quanto áreas podem ser subdivididas em três categorias gerais (LEVINE, 1996):

- Medidas de distribuição espacial: descrevem o centro, a dispersão, direção e forma da distribuição de uma variável;

- Medidas de autocorrelação espacial: descrevem a relação entre as diferentes localizações para uma variável simples, indicando o grau de concentração ou dispersão (por exemplo, análise de agrupamentos);

Medidas de distribuição espacial entre duas ou mais variáveis descrevem a correlação ou associação entre variáveis distribuídas no espaço, por exemplo, a correlação entre a localização de lojas de bebidas com pontos onde ocorrem muitos acidentes de trânsito.

Medidas de autocorrelação espacial surgem sempre que o valor de uma variável em um lugar do espaço está relacionado com seu valor em outros lugares no espaço. Nessa situação, observações separadas no espaço por certa distância espacial possuem valores similares (correlação).

O objetivo da estatística é medir o grau de associação espacial entre as observações de uma ou mais variáveis. Os índices analisados neste trabalho se dividem em duas categorias de abrangência, locais e globais. Quando o índice é global significa que pequenas diferenças locais perdem valor frente a grande diferenças globais. Índices locais geram resultados individualizados e possuem maior capacidade de representar alterações sutis que existam entre transições espaciais.

As equações a seguir apresentam os índices utilizados nesse trabalho para descrever textura em termos de autocorrelação espacial e distribuição espacial. Em todas as equações, i representa o ponto de referência e j o ponto analisado, com valor representado por x. A variável N representa a quantidade de pontos na análise, w é uma matriz de pertinência calculada pelo inverso da distância entre os pontos i e j, d_{ij} a distância entre os pontos i e j. Nas versões locais dos índices, o ponto de referência i é tomado uma vez para cada tom de cinza presente na amostra de análise. Sua obtenção, nesse caso, será detalhada durante a apresentação da metodologia.

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 (ANSELIN, 2001). Sua versão local é obtida pela seguinte equação:

$$I = \frac{x_i - \overline{x}}{\frac{\sum {x_j}^2}{N-1} - \overline{x}} \sum_{j=1}^{N} w_{ij}(x_j - \overline{x}), i \neq j \tag{2.21}$$

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

A função K de Ripley (RIPLEY, 1977) é um método de análise de segunda ordem comumente utilizada em análise de dados espaciais. Nos últimos trinta anos, sua aplicação foi utilizada nas mais diversas áreas como, por exemplo, geologia, epidemiologia, geomorfologia, criminologia (LANCASTER; DOWNES, 2004). Essa função pode ser utilizada para resumir um padrão de pontos, testar hipóteses sobre o padrão, estimar parâmetros e ajustar modelos:

$$R\left(d,i\right)=\sqrt{\frac{Ak(i,j)}{N}}, i \neq j, \tag{2.22}$$

onde d representa a função de distância utilizada na análise e também determina o parâmetro A, que representa a área da região do estudo (calculado a partir do ponto de referência i sob a distância d) e k é a função de pertinência que verifica se j está dentro do conjunto de análise determinado por i e d.

A estatística *JointCount* (ANSELIN, 2001) mede o número de bordas existentes entre as regiões espaciais. Por se tratar de uma medida de regiões, os padrões de bordas existentes são configurações binárias de como uma borda pode estar organizada em relação à vizinha direta. Sumariamente existem três padrões utilizados, através da combinação de presença determinado por preto e ausência determinando por branco: branco-branco (WW), preto-preto (BB), branco-preto (BW), definidos por:

$$\begin{aligned} BB &= \tfrac{0.5}{N}\sum\sum w_{ij}x_i x_j \\ BW &= \tfrac{0.5}{N}\sum\sum w_{ij}(x_i - x_j)^2 \\ WW &= \tfrac{0.5}{N}\sum\sum w_{ij}(1-x_i)(1-x_j) \end{aligned} \tag{2.23}$$

onde BB significa que em uma determinada ausência aconteceu repetidamente a distância 1. O contrário pode ser afirmado do padrão BB e o caso misto com BW. Existe autocorrelação positiva quando as informações estiverem agrupadas. Autocorrelação negativa quando estiverem formando um padrão linear, todavia

não agrupado e nenhuma autocorrelação, quando os pontos simplesmente estiverem organizados de maneira a não formar um padrão identificável.

A análise de vizinhança mais próxima (ANSELIN, 2001) (*Nearest Neighbor Analysis*) mede, a partir de um ponto de referência, a distância média de um ponto semelhante. Esse dado visa informar se a informação se encontra espacialmente homogênea ou heterogênea, sendo definido por:

$$\overline{d} = \frac{\sum_{i=1}^{N} \min(d_{ij})}{N}, i \neq j \tag{2.24}$$

onde d_{ij} representa a distância euclidiana do ponto i a j e N a quantidade total de pontos analisados.

No caso da distância média obtida, gerar um valor alto significa que existem muito dados que são diferentes agrupados em um pequeno espaço, provendo características de heterogeneidade.

2.5.4 Análise Geométrica

A análise geométrica é o estudo geométrico das formas de objetos. Especificamente um conjunto de medidas, que em sua essência caracterizam forma (circular, por exemplo) e aparência dos objetos de uma imagem através do comportamento do contorno ou medidas extraídas através da área do objeto em estudo.

Os resultados obtidos pela caracterização dependem muito da qualidade da imagem original. Ruídos, buracos ou qualquer outro defeito podem influenciar no método de descrição e por sua vez gerar resultados incoerentes. Dentre as várias medidas que descrevem formas geométricas, algumas são bem definidas e vastamente utilizadas na literatura. Parte-se no pressuposto que os achados clínicos da mamografia tenham comportamento circular sabendo que regiões segmentadas a partir de técnicas de detecção perdem a maior parte das imperfeições do contorno dado a natureza e densidade associada a núcleo único que define a região de interesse. Assim, as medidas utilizadas nesse trabalho medem aspectos de distribuição de densidade em forma circular.

As medidas são: Excentricidade, Circularidade, Compacidade, Solidicidade, Orientação (MONTERO; BRIBIESCA, 2009); Desproporção Circular e Densidade Circular (SOUSA *et al.*, 2010) e ainda propomos neste trabalho as seguintes

medidas: Densidade Quadrangular, Densidade Anular, Densidade Quadrática. Segue abaixo um detalhamento das medidas:

- Excentricidade: $E = \frac{minAxis}{maxAxis}$
  - Mede a desproporção da imagem ao longo do maior ($maxAxis$) e menor eixo ($mixAxis$);
- Circularidade: $C = \frac{4\pi A}{P^2}$
  - Mede o quão circular um objeto é em relação a seu perímetro (P) baseado em sua área (A);
- Compacidade: $C_o = \frac{P^2}{4\pi A}$
  - Mede o quão compacto está distribuído a área (A) de um objeto ao longo do perímetro (P);
- Desproporção Circular: $D = \frac{P}{2\pi Re}$
  - Mede o quanto a região analisada é proporcional a um circulo envolvente de raio Re e perímetro (P);
- Densidade Circular: $D = \frac{100n}{A}$
  - Mede qual a proporção da área (dado pela quantidade de pontos n) do objeto está dentro de um círculo envolvente de área A;
- Solidez: $S = \frac{A}{Aconvexa}$
  - Mede a distribuição da área do objeto (A) sobre sua área convexa ($Aconvexa$) verificando se existem buracos em sua borda;
- Densidade Quadrangular: $Dqi = \frac{A_{i\_quadrante}}{A_{quadrante}}$
  - Divide a região em quadrantes onde a área do objeto em cada quadrante ($A_{i\_quadrante}$) será comparado com a área total $A_{quadrante}$ do bounding box que envolve o objeto. O objetivo é indicar com está distribuído a área do objeto;
- Densidade Anular: $Da = \frac{A_{i\_anel}}{A_{anel}}$

  - O mesmo objetivo da Densidade Quadrangular, mas dividindo a região em anéis concêntricos; e

- Densidade Quadrática: $E = \frac{A}{A_{bb}}$
  - Mede a relação de proporção da área do objeto (A) com a área de um bounding box envolvente (A_{bb}).

Todas as medidas são baseadas em informações básicas de perímetro e área do objeto em estudo. O perímetro pode ser encontrado de uma maneira simples utilizando o fecho convexo de um objeto. A área pode ser obtida através da contagem de pixels pertencentes ao objeto.

Contudo, as regiões que são analisadas através da geometria naturalmente passam por problemas com a maneira que o perímetro convexo é calculado. Primeiro, em se tratando de achados mamográficos, é comum que estes possuam espículos ou protuberâncias. E ainda podemos acrescentar o fato do objeto em análise poder ter regiões internas vazias. Esses problemas não são tratados por algoritmos de cálculo de fecho convexo. Assim a distribuição da área do objeto sobre o perímetro convexo pode incorrer em erros.

Para minimizar os efeitos causados pelos problemas anteriores, utilizamos neste trabalho um tipo de geometria que se adapta aos pontos individuais do objeto e não a forma como um todo. A Geometria Côncava é capaz de produzir resultados mais aprimorados para o cálculo de delineamento de formas sendo este o assunto tratado na seção seguinte.

2.5.5 Geometria Côncava (*Alpha-Shapes*)

Assumindo que existe um certo conjunto $S \subset \mathbb{R}^d$ de n pontos num espaço d dimensional, pretendemos computar a forma dos n pontos. Um exemplo dessa forma é apresentado na Figura 2.13.

A partir desta abstração é possível obter a principal característica da geometria côncava ou α-shapes. O objetivo é que seja traçado um contorno através dos pontos que corresponda a forma côncava destes. O parâmetro que controla o quão fino ou granular é essa seleção é o α. Pode-se criar uma abstração onde ao redor de cada ponto é traçado uma circunferência de raio α, onde o ponto em análise

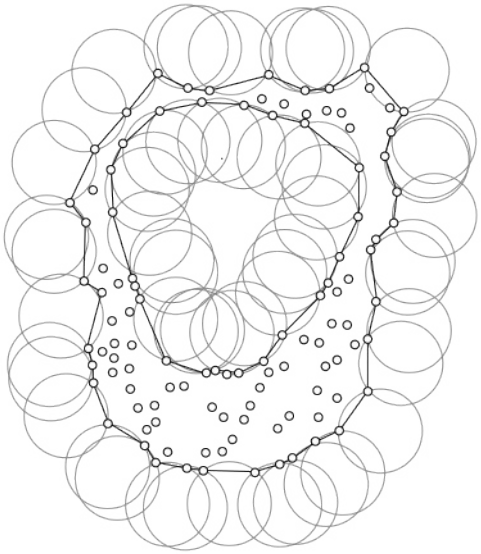


Figura 2.13: Contorno apresentado pela linha mais escura de uma coleção de pontos.

fica sempre no contorno da circunferência. O tamanho da circunferência poderia englobar apenas um ponto, neste caso $\alpha \to 0$ e por fim o contorno da forma é dado pelos próprios pontos. Na outra extremidade na analogia, temos $\alpha \to \infty$, ou seja, qual o valor máximo de circunferência em que não deixe buracos internos ou contornos mais longos. Neste último caso, o contorno será o próprio fecho convexo da coleção de pontos. Assim o parâmetro α é um regulador de contorno que sumariza o quão preciso deseja-se o contorno côncavo (MUCKE, 1994).

Formalmente, para $0 < \lambda < \infty$ tome uma λ-circunferência de raio λ. Defina-se 0-circunferência como um ponto e uma ∞-circunferência como um espaço aberto. Uma certa circunferência b é chamada de vazia se $b \cap S = \emptyset$. Assim, um k-simplexo $\triangle T$ é chamado de α-exposto se existe uma α -circunferência vazia onde $\triangle T = \partial b \cap S$ e ∂b é o fecho convexo de k pontos.

A Figura 2.14 exemplifica os conceitos de α-exposto. A partir da definição de α-exposto e de k-simplexo, define-se ∂S_α de um α-*shape* de um conjunto de pontos S, como todos os k-simplexos de S que $0 \leq k < d$ tal que sejam α-expostos:

$$\partial S_\alpha = \triangle T | T \subset S, |T| \leq d \; e \; \triangle T \; seja \; \alpha - exposto \tag{2.25}$$

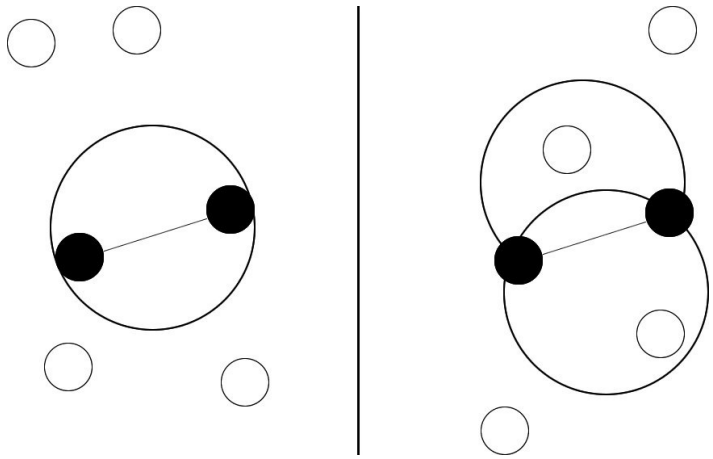


Figura 2.14: À esquerda, um k-simplexo α-exposto. À direita, um k-simplexo que não está α-exposto.

onde d representa a dimensão em que se encontra o conjunto de pontos. Embora ∂S_α possa representar pela notação um conjunto representante do contorno, em α-*shape* essa notação não significa somente isto já que podem existir vários contornos pertencentes ao mesmo α-*shape*. Um α-*shape* portanto é uma coleção de pontos que simplificam a superfície conectada.

Para simplificação dos cálculos de α-*shape*, assume-se que qualquer ∂S_α, para qualquer valor de α, pode ser obtido como um subconjunto da triangulação de Delaunay. Assim, dado um conjunto $S \subset \mathbb{R}^d$, a triangulação de Delaunay de S é o complexo $DT(S)$ consistindo de:

1. Todos os d-simplexos $\triangle T$ em que $T \subset S$ tal que a circunferência de T não contêm mais nenhum ponto de S, e

2. Todos os k-simplexos que sejam faces para outros simplexos em $DT(S)$.

Logo, se $\triangle T$ é um α-exposto simplexo de S, então $\triangle T \in DT(S)$. Assim, para obter um α-*shape* de um conjunto de pontos a partir de um conjunto de triângulos em uma DT(S) utilizando o algoritmo proposto em (MUCKE, 1994) é necessário para todo triângulo em $DT(S)$ fazer a seguinte verificação:

1. Se a circunferência que engloba o triângulo, tem raio menor que α e é vazia (não contém pontos internos); ou

2. Se $\triangle T$ é face para outro simplexos no conjunto α-complexo composto por todos simplexos válidos, representado por C_α

Simplexos que satisfazem uma das duas condições fazem parte do contorno e consequentemente do α-shape. A Figura 2.15 ilustra a obtenção de um α-complexo a partir de um triangulação de Delaunay. Verifica-se que somente faces que sejam menores que um círculo envolvente de tamanho α e que não possuam todas as faces vizinhas válidas permanecem como o conjunto final do α-shape.

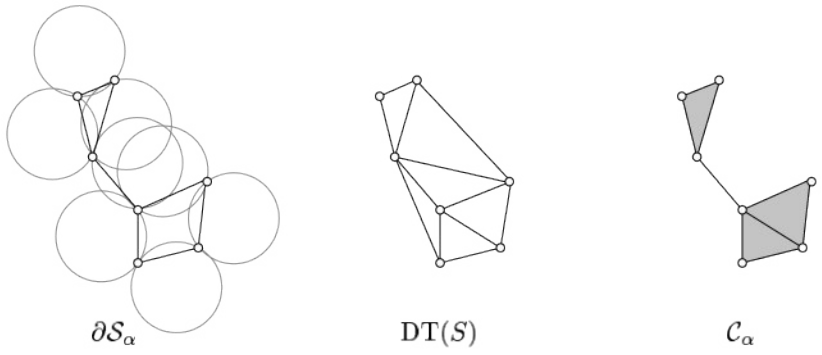


Figura 2.15: Representação de um C_α a partir de uma triangulação de Delaunay $DT(S)$. É possível verificar que somente simplexos que obedecem o tamanho máximo α ou que são contornos permanecem no C_α. Aqueles que não possuem todos os simplexos vizinhos no C_α também comporão o $\alpha-$shapes.

2.6 Reconhecimento de Padrões e Aprendizado de Máquina

Técnicas de Reconhecimento de Padrão são usadas para classificar ou descrever padrões ou objetos através de um conjunto de propriedades ou características extraídas. Um padrão é tudo aquilo para o qual existe uma entidade nomeável representante, geralmente, criada através do conhecimento cultural humano (BISHOP, 2006).

O reconhecimento de Padrão envolve dois processos: classificação, onde uma amostra de uma população qualquer é particionada em grupos chamados

classes; e reconhecimento, onde uma amostra desconhecida da mesma população é reconhecida como pertencente a uma das classes criadas. A classificação pode ser feita de duas formas: supervisionada e não supervisionada.

No processo de classificação não supervisionada, é examinado um conjunto de representantes de uma população. Esse conjunto é dividido em sub-conjuntos (classes) de acordo com critérios de similaridade intra-classe e dissimilaridade extra-classe. Esse processo também é chamado de agrupamento ou auto-organizável. Por outro lado, uma máquina para reconhecimento pode ser treinada previamente para identificar a classe de qualquer objeto desconhecido da mesma população. O processo de treinamento de um classificador é chamado de aprendizagem supervisionada.

Os objetos podem ser reconhecidos como pertencentes a uma determinada classe através de características capazes de realizar a distinção entre diferentes classes. Um número fixo de propriedades é usado para toda população e o conjunto de seus valores determina se um objeto pertence a uma classe ou não. As propriedades individuais são chamadas de características da população. Se existirem N características observáveis de uma população, forma-se um vetor de características. Logo, os vetores de características representam os objetos em uma população de objetos. O reconhecimento de padrão é realizado através de vetores de características.

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

Por fim, com o classificador devidamente treinado, é possível fazer o reconhecimento do padrão de um objeto que inicialmente pertence à mesma população, mas que é completamente desconhecido do classificador no processo de treinamento. A técnica atribui um rótulo a cada objeto, a partir do

conhecimento prévio obtido na etapa de treinamento, mesmo que o objeto não pertença a nenhuma das classes. Por isso se faz necessário que a tentativa de reconhecimento de padrão de um objeto seja realizada sobre aqueles do mesmo tipo se comparados aos de treinamento. Assim os padrões gerados na etapa de treinamento continuarão válidos na etapa de teste.

Este trabalho usa Máquina de Vetores de Suporte para realizar o reconhecimento de padrão de tecidos da mama extraídos de mamografias digitalizadas e descritos conforme a seção anterior.

2.6.1 Máquina de Vetores de Suporte

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 ideia básica por trás da MVS é construir um hiperplano como uma superfície de decisão, de tal maneira que a margem de separação entre as classes seja máxima. A margem de separação representa uma fonteira de decisão de classificação. Quanto maior a distância entre as classes (maior a fronteira), mais eficiente será o classificador ao tratar situações não previstas no treinamento.

O objetivo do treinamento através de MVS é a obtenção da superfície de decisão otimizando os limites de generalização, tornando-se eficiente para qualquer conjunto de dados que tenha sido descrito sobre o mesmo processo dos dados utilizados no treinamento.

A MVS é consideradas sistemas de aprendizagem que utilizam um espaço de hipóteses de funções lineares em um espaço de muitas dimensões. O método possui relação próxima à teoria de Redes Neurais Artificiais, percebida através do mapeamento de modelos através de funções complexas. Os algoritmos de treinamento das MVS possuem forte influência da teoria de otimização e de aprendizagem estatística. Em poucos anos, as MVS vêm demonstrando sua superioridade frente a outros classificadores em uma grande variedade de aplicações (CRISTIANINI; SHAWE-TAYLOR, 2000).

Em casos em que o conjunto de amostras é composto por duas classes separáveis, um classificador MVS é capaz de encontrar um hiperplano baseado em um conjunto de pontos denominados “vetores de suporte”, o qual maximiza a

margem de separação entre as classes. Por hiperplano entende-se uma superfície de separação de duas regiões num espaço multidimensional, onde o número de dimensões possíveis pode ser inclusive infinito. Mesmo quando as duas classes não são separáveis, a MVS é capaz de encontrar um hiperplano através do uso de conceitos pertencentes a teoria da otimização. O hiperplano ótimo (linha central), não somente separa as duas classes, mas possui a maior distância possível com relação aos vetores de suporte de cada classe.

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

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

$$y_i (w.x_i + b) \geq 1 \tag{2.26}$$

onde w é o vetor normal ao hiperplano de decisão e b o corte ou distância da função f em relação à origem. Os valores ótimos de w e b são obtidos ao minimizar a Equação 2.27, de acordo com a restrição dada pela Equação 2.26 (BISHOP, 2006):

$$\phi(w) = \frac{w^2}{2} \tag{2.27}$$

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 2.26 sejam obtidas.

O problema de otimização passa a ser então a minimização da Equação 2.27, de acordo com a restrição imposta pela Equação 2.26, fazendo com que C seja um parâmetro de treinamento que estabelece um equilíbrio entre a complexidade do modelo e o erro de treinamento que deve ser selecionado pelo usuário. Assim, define-se o seguinte sistema:

$$\phi(w, \xi) = \frac{w^2}{2} + C \sum_{i=1}^{N} \xi_i \tag{2.28}$$

para:

$$y_i\left((w.x_i)+b\right)+\xi_i \geq 1 \tag{2.29}$$

Usando a teoria dos multiplicadores de Lagrange, o objetivo passa a ser encontrar os multiplicadores de Lagrange α_i ótimos que satisfazem a Equação 2.30, dada por:

$$w\left(\propto\right)=\sum_{i=1}^{N}\propto_i-\frac{1}{2}\sum_{i,j=1}^{N}\propto_i\propto_j x_i x_j \phi(x_i x_j) \tag{2.30}$$

sujeito a:

$$\sum_{i=1}^{N}\alpha_i x_i=0, \quad 0\leq\alpha_i\leq C \tag{2.31}$$

Apenas os pontos onde a restrição presente na Equação 2.26 seja exatamente igual à unidade têm os α correspondentes diferentes de zero. Esses pontos são chamados de vetores de suporte, pois se localizam geometricamente sobre as margens. Tais pontos possuem fundamental importância na definição do hiperplano ótimo, pois os mesmos delimitam a margem do conjunto de treinamento.

A Figura 2.16 destaca os pontos que representam os vetores de suporte. Os pontos além da margem não influenciam decisivamente na determinação do hiperplano, enquanto que os vetores de suporte, por terem pesos não nulos, são decisivos.

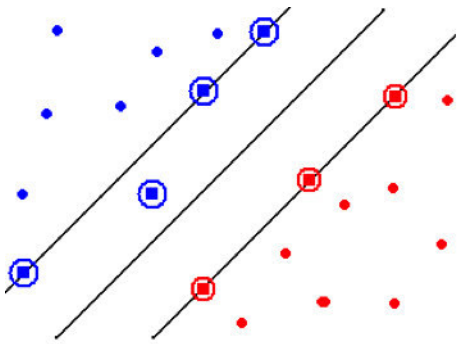


Figura 2.16: Vetores de suporte destacados por círculos.

Para que a MVS possa classificar amostras que não são linearmente separáveis, é necessário uma transformação não-linear que transforme o espaço de entrada

(dados) para um novo espaço (espaço de características). Esse espaço deve apresentar dimensão suficientemente grande, e através dele, a amostra pode ser linearmente separável. Dessa maneira, o hiperplano de separação é definido como uma função linear de vetores retirados do espaço de características ao invés do espaço de entrada original. Essa construção depende da utilização de uma função Kernel (HAYKIN; ENGEL, 2001) que pode realizar o mapeamento das amostras para um espaço de dimensão mais elevada sem aumentar a complexidade dos cálculos.

Substituindo a Equação 2.30 pela equação a seguir com a utilização do kernel, tem-se:

$$w(\propto) = \sum_{i=1}^{N} \propto_i - \frac{1}{2} \sum_{i,j=1}^{N} \propto_i \propto_j y_i y_j K(x_i x_j) \tag{2.32}$$

Uma importante família de funções de kernel 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 por:

$$R(x_i, x_j) = \epsilon^{-\gamma(x_i - x_j)^2} \tag{2.33}$$

onde γ é um parâmetro informado externamente que representa o fator regulador de complexidade da função de mapeamento dos dados, tornando possível a adaptação dos modelos de classificação a qualquer ordem de complexidade dos dados.

2.6.2 Validação de Resultados

Após o reconhecimento é necessário realizar sua avaliação para verificar se os resultados obtidos foram satisfatórios e localizar onde são necessárias melhorias. Tal avaliação envolve a comparação de medidas obtidas simultaneamente, utilizando o teste em estudo e um teste de referência. Os estudos de avaliação implicam que esse teste de referência seja o apropriado. Um dos grandes problemas inerentes a este tipo de estudo é o fato de, por vezes, não existir uma referência, usando-se, então, o melhor procedimento disponível como procedimento de referência. É importante frisar que uma medida é válida se provém de um procedimento válido.

O valor clínico de um teste está relacionado com a sua especificidade e sensibilidade (ALTMAN; BLAND, 1994). Ele deve fornecer uma boa indicação preliminar de quais indivíduos têm a doença e quais não têm, e isto só se consegue se os métodos utilizados forem válidos.

A sensibilidade (S) é a proporção de indivíduos doentes (VP) que possuem um teste positivo, isto é, a probabilidade de estando doente, um indivíduo ter um teste positivo (percentagem de vezes que o teste acerta). A especificidade (E) é a proporção de indivíduos não doentes (VN) que possuem um teste negativo ou a probabilidade de, não estando doente, ter um teste negativo. A sensibilidade define-se, então, como sendo a capacidade de um teste para identificar corretamente aqueles indivíduos que possuem uma determinada doença, enquanto que a especificidade é definida como a capacidade do teste para identificar corretamente aqueles que não a possuem. Ambas são determinadas pela comparação dos resultados obtidos num determinado teste com os resultados de métodos de diagnóstico mais seguros (de referência). A extensão em que os resultados de um teste coincidem com o de referência dá uma medida da sensibilidade e especificidade desse teste.

Quando indivíduos doentes são considerados negativos ou normais, os respectivos resultados deste teste são chamados "falsos negativos" (FN). Por outro lado, quando indivíduos não doentes são considerados como doentes, os resultados deste teste são denominados "falsos positivos" (FP). Note-se que a percentagem de falsos negativos é o complemento da sensibilidade e a percentagem de falsos positivos é o complemento da especificidade. Quando a sensibilidade é de 100%, temos a certeza que o teste nunca se engana nos falsos negativos. A taxa média de falso positivos detectados é uma importante medida de avaliação de acurácia em relação a proporção de verdadeiros positivos encontrados. A medida indica a taxa de precisão mesmo com a prevalência de alta sensibilidade. A acurácia (Acc) total do diagnóstico é portanto obtida como a taxa de acerto total do método. Sensibilidade, especificidade e acurácia são definidas pelas Equações 2.34, 2.35 e 2.36.

$$S = \frac{VP}{VP + FN} \tag{2.34}$$

$$E = \frac{VN}{VN + FP} \tag{2.35}$$

$$Acc = \frac{VP + VN}{VP + VN + FN + FP} \tag{2.36}$$

Além de estatísticas para diagnóstico clínico, também é necessário especificar testes para avaliação da etapa interna de detecção de achados clínicos. Para tanto, as estatísticas de proporção de falso positivos e negativos por exame é de fundamental importância para mensurar o quão correto se encontra o método para encontrar achados clínicos ao mesmo tempo em que gera poucos suspeitos.

Para essa etapa, a curva ROC (*Receiver Operating Characteristic*) (OBUCHOWSKI, 2005) não é capaz de representar corretamente a informação de precisão de acerto já que se faz necessário a localização dos achados. Em uma situação comum, um exame pode gerar um certo número de falso positivos enquanto que outros cinco exames não geram qualquer falso positivo. A técnicas mais eficiente para representar o quão bom é o método para detectar achados clínicos numa imagem é um extensão da curva ROC chamada de FROC (*Free Receiver Operating Characteristic*) (GUR *et al.*, 2009).

A rotina de construção da curva FROC envolve o problema de que um especialista ao começar a examinar uma determinada mamografia, ainda desconhece a localização da lesão (se é que realmente existe) e quantas são (BORNEFALK; HERMANSSON, 2005). Seu trabalho consiste então em relacionar as regiões suspeitas e prover a cada uma um grau de confiança quanto ao fato de ser realmente uma lesão. Assim, define-se para a construção de uma curva FROC as estatísticas:

1. LL = lesão localizada corretamente;
2. NL = lesão erroneamente localizada (não existente);
3. LLF = fração de lesões corretamente localizadas (LL / número de lesões); e
4. NLF = fração de lesões erroneamente localizadas (NL / número de imagens).

onde $0 \leq LLF \leq 1$ e $0 \leq NLF$.

A curva FROC é o resultado do gráfico de LLF por NLF por imagem, cumulativamente, conforme exemplo apresentado na Figura 2.17. Perceba que a curva pode crescer indefinidamente para a direita, mas o eixo das ordenadas é limitado ao valor 1. Após a construção da curva é possível verificar o comportamento da composição da taxa de verdadeiros positivos (LLF) e falso positivos (NLF) detectados por imagem.

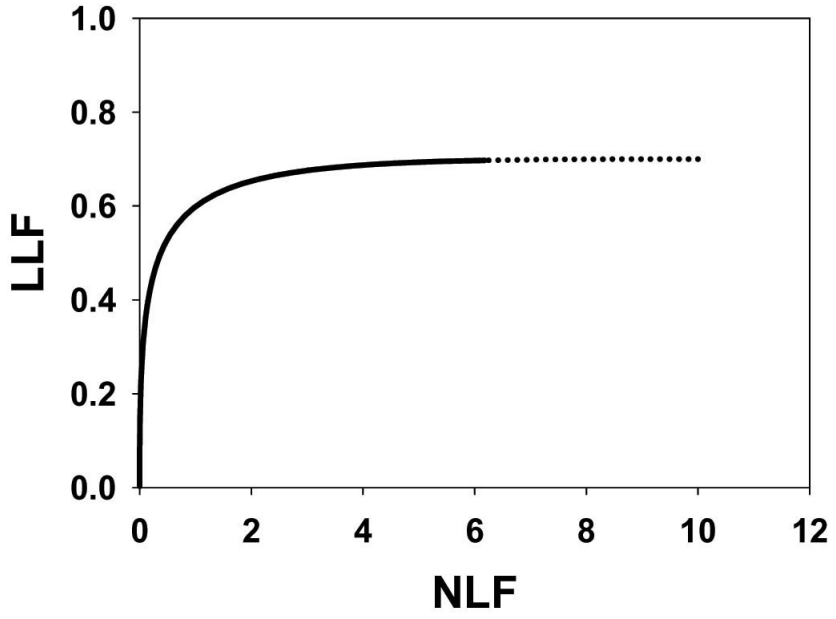


Figura 2.17: Um exemplo curva FROC. Verifique que o eixo de falso positivos pode crescer indefinidamente enquanto que a taxa de acerto de verdadeiro positivos varia entre 0 a 1. Fonte: (CHAKRABORTY, 2014)

Capítulo 3

Metodologia Proposta

Este capítulo apresenta a metodologia proposta para a detecção automatizada de massas através de imagens de mamografias. O esquema geral desta metodologia é apresentado na Figura 3.1.

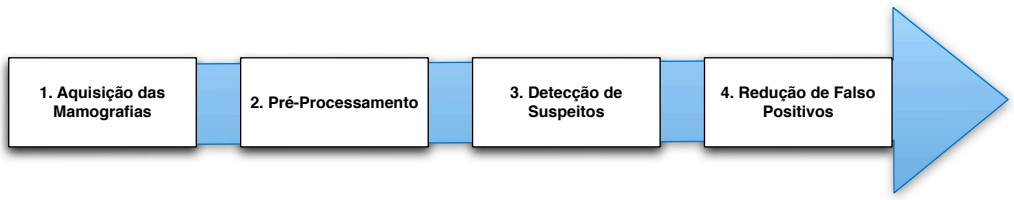


Figura 3.1: Etapas da metodologia proposta para detecção automatizada de massas através de imagens de mamografia.

A metodologia como um todo consiste na união de etapas, cada uma com função específica. A primeira etapa consiste na aquisição da base de imagens com especificação da localização do achado clínico para posterior avaliação de resultado. A segunda etapa consiste no realce da imagem de mamografia, tratando de minimizar efeitos de ruídos, retirar informações de fundo, retirar o músculo peitoral na projeção MLO e aumentar o contraste de estruturas suspeitas na imagem.

A terceira etapa consiste em obter as possíveis regiões suspeitas que possuem semelhança com uma massa. Como normalmente muitas regiões suspeitas são na verdade falso positivos, a quarta etapa tem como objetivo apenas a redução

de falso positivos ao mesmo tempo que mantém alta taxa de acerto para os verdadeiros positivos.

Esta metodologia propõe uma abordagem automatizada de detecção de massas. Logo, para que seja usada, alguns parâmetros devem ser estimados e estudados para a base de imagens. Todavia, é um dos objetivos desta pesquisa prover mecanismos de adaptação automática dos parâmetros usados. As seções seguintes têm como objetivo apresentar mais especificamente os procedimentos realizados durante a metodologia.

3.1 Aquisição das mamografias

As mamografias utilizadas para teste e validação da metodologia proposta foram obtidas através da base mini-MIAS (SUCKLING *et al.*, 1994) e DDSM (*Digital Database for Screening Mamography*) (HEATH *et al.*, 1998).

A base mini-MIAS contém as mamografias das mamas esquerda e direita de 161 pacientes de idades entre 50 a 65 anos. Todas as imagens foram digitalizadas na resolução de 1024 x 1024, com 8 bits de profundidade por pixel e foram obtidas na projeção Médio-Lateral Oblíqua (MLO). Todas as imagens possuem um arquivo descritivo para informar a presença ou ausência de lesão, sua localização, o tipo de lesão quando existir (massa, microcalcificações, distorção arquitetural), a caracterização da lesão quanto a malignidade e a densidade da mama (*fatty, glandular e dense*)

A base DDSM possui 2620 casos adquiridos através das seguintes instituições americanas: *Massachusetts General Hospital*, *Wake Forest University*, e *Washington University in St. Louis School of Medicine*. Cada caso contém quatro mamografias referentes as imagens da mama esquerda e direita nas visões Médio-Lateral Oblíqua (MLO) e Crânio Caudal (CC). Os dados são constituídos de estudos de pacientes de diferentes origens étnicas e raciais. A base ainda disponibiliza informações sobre a paciente, tal como a idade e a densidade da mama. Também são informados o tipo do digitalizador utilizado e a resolução de cada imagem. Imagens com áreas suspeitas possuem a descrição da anormalidade, o diagnóstico e a localização da mesma na imagem. As descrições de lesões em imagens mamográficas seguem os termos lexicográficos publicados no BI-

RADS. Estes achados clínicos foram previamente segmentados por métodos computacionais e, em uma etapa seguinte, aprimorados através de especialistas.

3.2 Pré-Processamento

A etapa de pré-processamento consiste em melhorar feições presentes na imagem digitalizada de mamografia que em geral apresenta baixo contraste. Os objetivos nesta etapa são:

- Remoção do fundo da imagem;
- Remoção do músculo peitoral; e
- Aumento de contraste e ênfase de regiões suspeitas.

Um exemplo de imagem de mamografia na incidência MLO é apresentado na Figura 3.2 (a). Todas as imagens, independente da base utilizada, são redimensionadas para altura 1024, com redimensionamento proporcional de largura. O redimensionamento é realizado para manter constantes parâmetros dependentes de resolução e aumentar a velocidade de processamento dos métodos sem prejuízo para seu desempenho.

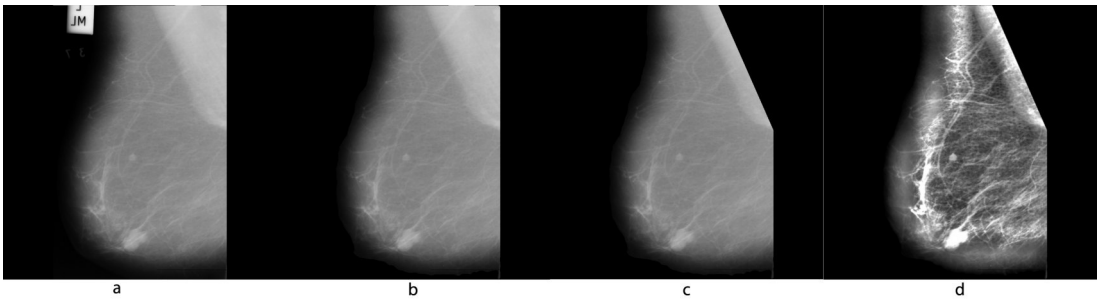


Figura 3.2: Passos do Pré-Processamento: (a) a imagem original, (b) após retirada do fundo, (c) após retirada do músculo peitoral e (d) após melhoramento com CLAHE e *Constrast Streching*.

A primeira tarefa consiste na retirada do fundo da imagem e consequentemente da anotação do paciente. Para realizar essa etapa, aplica-se o Kmeans (MACQUEEN *et al.*, 1967) com 2 grupos. Um grupo consiste no

fundo e o outro em regiões com intensidade maior que zero. Seleciona-se a partir do resultado da execução do Kmeans como resultado final o maior grupo de intensidade maior que zero (Figura 3.2 (b)).

Após a retirada do fundo, o próximo passo é retirar o músculo peitoral. Para tanto, utilizamos o algoritmo proposto em (SAMPAIO *et al.*, 2011) que inicialmente identifica o quadrante onde está o músculo peitoral. Em seguida é aplicado o filtro de Canny (CANNY, 1986) para obter contornos que são reduzidos usando uma operação de erosão morfológica. Contornos suficientemente pequenos e em direções contrárias ao músculo peitoral são eliminados. Finalmente, a transformada de Hough (GONZALEZ; WOODS, 2010) é usada para estimar o contorno que melhor define a borda do músculo peitoral e assim delimitar sua região.

O resultado é exemplificado pela Figura 3.2 (c). Por último, é realizado o realce de contraste utilizando os métodos CLAHE (Seção 2.3.1) e *Contrast Streching* (Seção 2.3.3), sendo o último para adaptação de faixas. Ainda é realizada uma suavização por média para redução de ruídos. Os parâmetros utilizados são adaptativamente calculados com base no contraste calculado para a imagem resultante da retirada dos artefatos de fundo e músculo peitoral (quando aplicável). O contraste é calculado usando as medidas de Haralick (HARALICK *et al.*, 1973), com matriz de co-ocorrência de direção zero graus e distância 1. O valor obtido de contraste é normalizado na faixa 0 a 1, para todas as imagens, com intuito de refletir proporcionalidade nos parâmetros calculados sobre ele. O valor mínimo de contraste calculado foi de 1,67 para todas as imagens. O valor máximo foi de 138.

O parâmetro α utilizado no CLAHE é calculado conforme:

$$\alpha = 0,007C + 0,011 \tag{3.1}$$

onde C representa o contraste calculado para a imagem, $0,011$ representa o α mínimo a ser aplicado na imagem que possuir pouca densidade. Em imagens de alta densidade, o valor de contraste aplicado deve ser de $0,018$ correspondendo ao valor máximo. Estes valores limites foram empiricamente calculados a partir de análises sobre as imagens da base. Os valores intermediários são adaptados linearmente dentro da faixa. Novas imagens, com contraste menor ou maior, são

adaptadas ao valor mínimo ou máximo de α, respectivamente.

Da mesma maneira que α, o tamanho da janela utilizada no CLAHE e na suavização por média é calculada segundo o contraste, seguindo a seguinte relação.

$$T_j = (8C + 6) \tag{3.2}$$

onde as constantes foram empiricamente calculadas para melhor se adaptarem à imagem. O tamanho de janela T_j é utilizado como padrão para todas as etapas da metodologia.

O resultado final do pré-processamento é exemplificado pela Figura 3.2 (d). É possível constatar que a estrutura da massa foi realçada e possui um contraste maior em relação ao fundo quase preto ao seu redor. A imagem final também retira o fundo, anotações e músculo peitoral que poderiam influenciar no resultado. Todavia, outras estruturas que não são massas também são realçadas no processo e são acompanhadas nas próximas etapas.

3.3 Detecção de Regiões Suspeitas

A detecção é tida como um dos processos mais dependentes da imagem e do que se está buscando nela. O grande desafio desta metodologia é buscar uma forma automatizada de lidar com diferentes tipos de imagens e gerar uma seleção de regiões suspeitas que sempre contenham a região da massa e que minimize a quantidade de falso positivos.

A suposição a ser utilizada para detecção automatizada é que as regiões de massas a serem detectadas como suspeitas compartilham características como: alta média de tonalidades, normalmente com forma definida (circulares ou espiculadas) e que normalmente não são regiões que ocupam uma grande área proporcional da imagem como um todo. Essa suposição será colocada como objetivo das subetapas da detecção apresentadas na Figura 3.3 e detalhadas nas próximas seções.

3.3.1 *MeanShift*

A partir da imagem realçada, é aplicado o agrupamento usando o algoritmo *MeanShift* (Seção 2.4.1) sob cada uma das imagens. Este é um método não

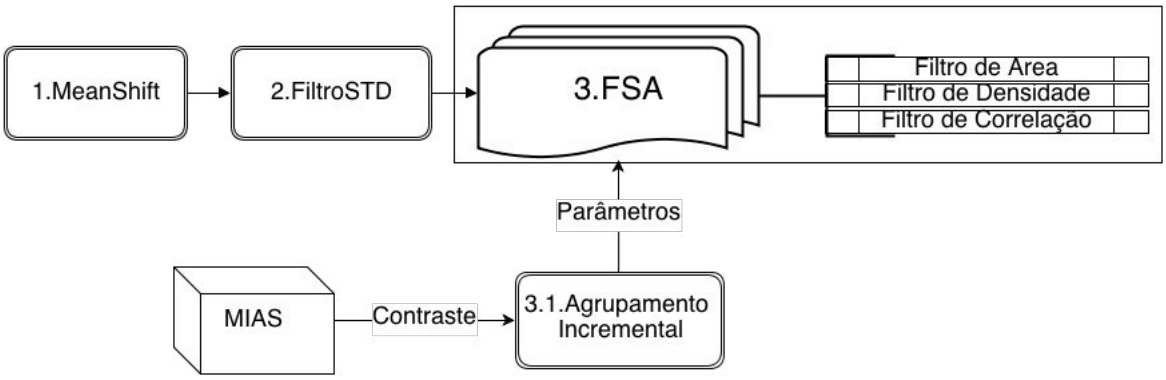


Figura 3.3: Passos utilizados pela etapa de detecção de regiões suspeitas

supervisionado de agrupamento no qual não é necessário a informação da quantidade de grupos a serem gerados. Seus parâmetros são o tamanho da janela na qual se deve estimar a função de densidade e a distância máxima entre tons para um mesmo grupo, ajustados para todas as imagens com os valores 15 e 60 respectivamente. Os valores foram empiricamente obtidos a partir da análise do objetivo da utilização do *MeanShift* nessa etapa que consiste no simples agrupamento de regiões de alta média de tonalidades.

Como não existe o controle da quantidade de grupos, todas a fronteiras de regiões, que obedeçam a restrição de distância máxima de tonalidades, serão separadas como grupos, criando uma superfície de nível onde o grupo de maior tonalidade fica no centro contornado por grupos de tonalidades sucessivamente inferiores. Embora não desejamos definir uma quantidade máxima de grupos, também não desejamos uma quantidade excessivamente grande. Logo, para suavizar os efeitos do agrupamento, aplicamos sobre a imagem resultante do *MeanShift* uma suavização por média (com janela de tamanho T_j) para provocar um leve borramento dos grupos.

Os resultados obtidos pela etapa *MeanShift* são apresentados na Figura 3.4. É possível verificar os agrupamentos, que esses se encontram bem definidos, mas que apresentam subdivisões internas.

Figura 3.4: Resultados obtidos do agrupamento usando *MeanShift* e filtro da média em (c) sobre as imagens após o realce de contraste (b).

3.3.2 Filtro Desvio Padrão (STD)

Mesmo com subdivisões internas, é evidente a geração de contornos mais definidos que separam regiões de alta tonalidades média de regiões de baixa tonalidade média. Com o intuito de separar esses grupos, aplicamos um Filtro de Desvio Padrão (STD) para retirar toda região de fronteira entre alta intensidade e baixa intensidade. O filtro basicamente calcula o desvio padrão dentro de uma janela configurada para o tamanho 3x3 (mínimo possível). A restrição imposta para separação de borda consiste em retirar todos os pixels que tenham ao seu redor um desvio padrão maior que 1. A Figura 3.5 apresenta o resultado obtido pelo filtro de desvio padrão.

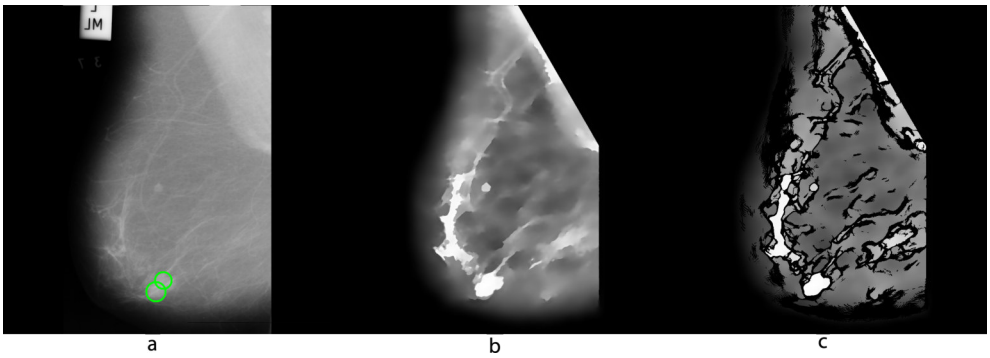


Figura 3.5: Resultado obtido após a aplicação do STD em (c) sobre a imagem após o *MeanShift* em (b).

3.3.3 Fast Scanning Algorithm (FSA)

Após o STD, é necessário separar as regiões suspeitas. Para tanto, esta metodologia utiliza uma extensão do método *Fast Scanning Algorithm* (FSA) (Seção 2.4.2). Incluímos três restrições conforme a suposição de que uma região de massa está entre as regiões de maior intensidade e mudamos a etapa de união original do algoritmo, realizada no final de sua execução. As extensões são:

1. Filtro de área: todas as regiões que possuam área menor que 200 ou maior que 15000 pixels são eliminadas. A eliminação de regiões com área menor que 200 representa uma proteção a ruídos que podem ter sido provocados por super segmentação. Áreas maiores que 15000 não representam massas e tipicamente são estruturas remanescentes do músculo peitoral, ou do fundo da imagem;

2. Filtro de densidade: consiste na identificação de regiões de maior densidade. A densidade é determinada pela média de tonalidades presente dentro da região. Quanto maior a média, maior a densidade. Inicialmente, a região de maior densidade presente é utilizada como molde para se identificar as demais regiões. Estas são identificadas a partir da pertinência de semelhança de densidade em relação a região de maior densidade. O parâmetro pertinência varia entre 0 e 1, onde o valor 1 representa maior similaridade; e

3. Filtro de correlação: consiste na identificação das regiões de maior correlação em relação a que possui a maior densidade dentre todas (selecionada no filtro anterior). A informação de correlação é obtida através do histograma (Seção 2.5.1) e varia entre 0 e 1, onde 1 representa correlação máxima.

As extensões são aplicadas em sequência, provocando uma redução gradativa de regiões suspeitas. A última etapa do FSA consiste em unir grupos identificados que possuam área muito pequena com outros grupos que estejam conectados. O STD provocaram uma mudança nessa etapa ao desconectarem regiões e retirarem regiões muito pequenas. Logo, propomos uma mudança no algoritmo original, como uma nova extensão, que consiste em unir regiões conectadas ou não, remanescentes dos filtros que possuam uma regra de correlação de textura satisfeita.

O objetivo é unir regiões de alta densidade, que representem tecido granular, e aumentar a capacidade descritiva dos algoritmos de extração utilizados na etapa seguinte, de redução de falso positivos, ao incluir mais informações nas regiões.

Para tanto, cada região possui uma informação calculada de centro de massa. A união é realizada sobre cada região remanescente levando em consideração a distância em relação ao centro para determinar conectividade e correlação dos histogramas e determinar a similaridade de textura. A distância máxima utilizada é igual a A/10, onde A representa a altura da imagem. A correlação mínima que deve existir é de 0,6 para representar regiões similares. A distância máxima entre grupos foi empiricamente determinada da observação que massas normalmente não ocupam mais que 20% da altura da imagem. O grau de similaridade foi empiricamente determinado para maximizar a junção entre regiões de densidade comum, situadas na localidade determinado pelo raio de atuação da distância máxima entre grupos.

Os outros dois parâmetros da FSA, pertinência de média e correlação, são obtidos por grupos de imagens que compartilham informação de contraste comum. Utilizamos um modelo de agrupamento incremental para a determinação destes parâmetros, descrito a seguir.

Modelo de Agrupamento Incremental para Seleção de Parâmetros usados no FSA

Verificamos que para a aplicação do FSA se faz necessário estimar os parâmetros utilizados nos filtros de média e correlação que consistem na pertinência em relação a região de maior média. A escolha de um único parâmetro para todas as imagens leva a resultados ruins haja visto que cada imagem possui um comportamento de densidade diferente. Todavia, a escolha de um parâmetro para cada imagem torna o método manual e inviável.

Então, propomos para seleção de parâmetros um modelo incremental de agrupamento das imagens tomando como base informação de contraste das mamografias. O contraste determina a densidade das imagens. Imagens de baixo contraste são também imagem de alta densidade, onde existe uma concentração muito grande de altas intensidades em grandes regiões. Imagens de alto contraste possuem baixa densidade e também regiões de alta intensidade concentradas em

pequenas regiões, normalmente isoladas. Imagens de contraste mediano possuem regiões de alta intensidade concentrada em inúmeras pequenas regiões distribuídas ao longo da imagem.

Assim como a adaptação dos parâmetros de pré-processamento, a informação de contraste, calculada através das medidas de Haralick *et al.* (1973) sobre a imagem resultante do filtro STD é usada como critério de adaptação no modelo de agrupamento. A partir da informação de contraste, o modelo incremental realiza os seguintes passos para agrupar o conjunto de imagens U em agrupamentos G_k. Um agrupamento possui um conjunto de imagens que apresentam um contraste semelhante e o valor médio do contraste das imagens presentes, utilizado como núcleo.

1. Uma imagem X_i é retirada da base de imagens a ser analisada U
2. Procurar um agrupamento G_k onde X_i possa ser inserido. Para tanto, a distância interna do grupo, quando inserida a imagem X_i deve ser menor que D_{max}.
   (a) Caso não exista um agrupamento viável: um novo é criado e a imagem X_i inserida.
   (b) Caso existam múltiplas opções, a imagem X_i é inserida no grupo que possui a menor distância em relação ao centro.
3. Ajustar agrupamento G_k modificado/criado calculando os valores do elemento representante.
4. Repetir os passos 1-3 até que todas as imagens tenham um agrupamento correspondente.

Com intuito de reduzir o impacto que a ordem que as imagens pode provocar na geração dos grupos, o processo é repetido iterativamente após o primeiro agrupamento afim de verificar uma troca de imagens entre grupos e fazer com o erro total do agrupamento seja minimizado.

Ao final, o melhor conjunto de parâmetros de pertinência de média e de correlação é obtido para cada grupo utilizando a marcação de cada massa provida inicialmente pela base como mecanismo de treinamento. Os melhores

parâmetros são aqueles que maximizam a sensibilidade, minimizam a taxa de falso positivos por imagem e possuem a menor quantidade de grupos possível. O número médio de falso positivos cresce quando a quantidade de grupos diminui, significando que com mais grupos existe uma maior capacidade do algoritmo de gerar parâmetros melhores. Todavia um grande número de grupos torna o processo de seleção de parâmetros superajustado. Assim, o número de grupos é adicionado como restrição para seleção de parâmetros e o conjunto de parâmetros pode automaticamente ser selecionado, sendo aquele que maximize a sensibilidade, minimize a taxa média de falso positivos e quantidade de grupos, conforme apresentado no pseudocódigo abaixo:

```
Parametros SelecionaIncremental(ListaParametros resultados) {
        ListaParametros C =
            selecionaMelhoresPorSensibilidade(resultados);
        ListaParametros selFinal;
        if (C.tamanho > 1)
                selFinal = selecionaMelhorPorGrupoEFP(C);
        else
                selFinal = C;

        return selFinal.primeiroElemento();
}
```

Uma nova imagem que não estava presente durante a etapa de treinamento (agrupamento) terá seus parâmetros escolhidos de acordo com o grupo que mais tenha similaridade. O processo pode ser repetido iterativamente para prover adaptação de parâmetros. A política de atualização pode ser definida externamente conforme a situação que melhor determine o comportamento desejado do mecanismos de agrupamento.

3.4 Redução de Falso Positivos

Após a detecção de áreas suspeitas, verifica-se que muitas não contêm de fato uma lesão. Assim, é muito importante a redução destes falso positivos.

Chama a atenção a necessidade de reduzir a quantidade de falso positivos, mantendo o acerto total nos verdadeiros positivos. Um erro na perda de um verdadeiro positivo é interpretado como mais prejudicial, pois envolve a ocorrência de uma lesão não detectada.

Para a redução dos falso positivos, utiliza-se medidas de textura (Seção 2.5.2 e 2.5.3) e geometria (Seção 2.5.4). Portanto, esta etapa usa análise de diversidade, análise espacial e geometria côncava para representar e descrever regiões extraídas de imagens de mamografia para em seguida classificá-las.

O esquema básico desta etapa é apresentado na Figura 3.6, onde basicamente a redução de falso positivos (RFP) é realizada em dois momentos, independente do tipo de função usada para obter características.

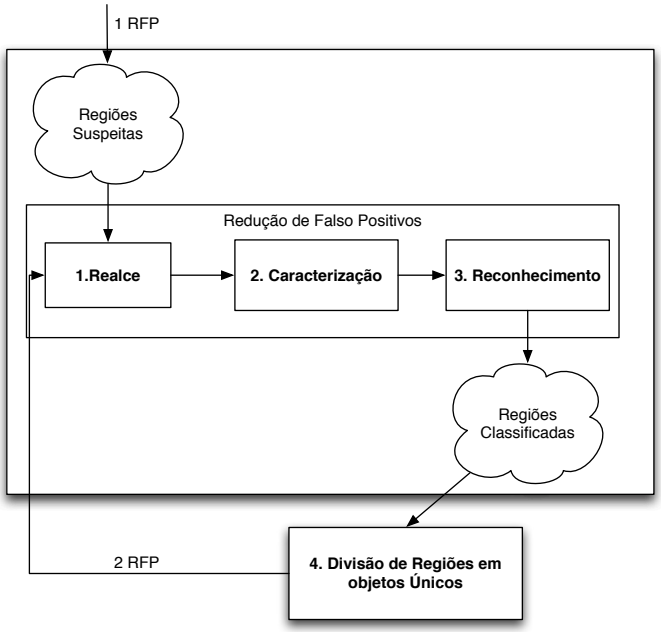


Figura 3.6: Fluxo de atividades para redução de falso positivos.

No primeiro momento, as regiões identificadas na etapa anterior são submetidas à caracterização e reconhecimento. Muitas das regiões podem não ser simples e conter vários objetos unidos devido a regra de união do algoritmo *Fast Scanning*.

Com o intuito de minimizar esse efeito, uma segunda etapa de reconhecimento é realizada usando a mesma metodologia de descrição sobre o conjunto de regiões

resultado do primeiro momento. Neste segundo momento, regiões que possuem múltiplos objetos são divididas em múltiplas regiões, com a classificação ajustada de acordo com a informação provida pela base. Em seguida, o processo de caracterização e reconhecimento.

As etapas internas da redução de falso positivos segue fluxos diferentes de acordo com o tipo de caracterização que é realizada, conforme apresentado na Figura 3.7. O detalhamento de cada etapa será tratado nas subseções seguintes.

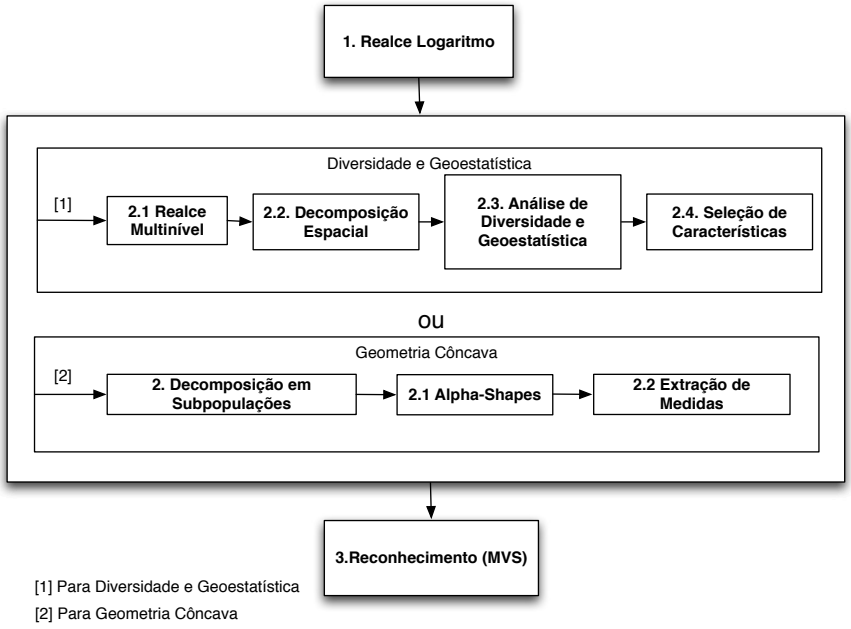


Figura 3.7: Etapas da descrição e classificação usando os diferentes índices estudados, verificando a divisão de fluxo de acordo com o tipo de característica extraída.

3.4.1 Realce Logarítmico e Multinível

As regiões suspeitas são submetidas inicialmente a um processo de realce. O objetivo desta etapa é aumentar a distinção das texturas que representam os padrões de regiões normais e de massas. Esta é realizada em duas partes.

A primeira consiste na aplicação do filtro de realce logaritmo (Seção 2.3.2). O objetivo do filtro é intensificar a participação de tons de cinza raros na distribuição

total da imagem.

A segunda parte do realce consiste na decomposição da imagem em outras seis imagens com faixa máxima de valores diferentes. Chamamos esta decomposição de multinível. A finalidade é gerar várias quantizações, calculadas de maneira linear, da mesma região e facilitar a codificação de padrões de tons de cinza presentes em agrupamentos diferentes. São utilizados como faixas: 256, 128, 64, 32, 16, 8.

Essa segunda etapa somente é realizada quando a análise se tratar de textura. Como será abordado em seções seguintes, a análise geométrica realiza um agrupamento de pontos que utiliza um mecanismo de quantização não linear. Cada região resultante do realce multinível, correspondente a cada quantização, é utilizada separadamente durante a etapa de descrição. Todavia, as cinco são utilizadas em conjunto durante a etapa de reconhecimento.

3.4.2 Decomposição Espacial - Zoneamento

Após a etapa de realce, as imagens resultantes passam por um processo de agrupamento de características locais com o objetivo de melhorar a etapa de descrição e consequentemente a etapa de reconhecimento. Essa análise é conduzida na forma da divisão espacial da região. O objetivo é quantificar os descritores da etapa seguinte associando-os a recortes delimitados dentro da área como um todo e assim preservar associações espaciais de vizinhança e concentração.

A região original é decomposta em recortes: retangular (Horizontal, Vertical, Janelas, Janelas Centralizadas), Circular, Anel e Diagonais.

A decomposição retangular é feita utilizando quatro padrões específicos, onde cada um representa uma divisão da região original:

- Horizontal: a região é subdividida em recortes horizontais de mesma área;
- Vertical: a região é subdivida em recortes verticais de mesma área;
- Janelas: união da Horizontal e Vertical, provocando a divisão da região em janelas de mesma área; e
- Janelas Centralizadas: alteração da divisão em janelas para preservar uma área maior central em detrimento das bordas.

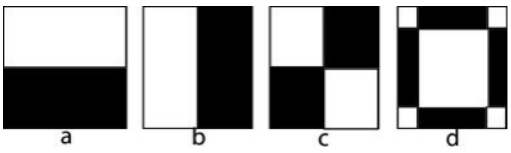


Figura 3.8: Divisões retangulares aplicadas à região para preservar associações locais antes da aplicação de métodos de descrição: (a) horizontal, (b) vertical, (c) janelas e em (d) janelas centralizadas.

A Figura 3.8 exemplifica cada uma das abordagens. A quantidade de recortes horizontais, verticais e de janelas são configuráveis. São utilizados quatro recortes horizontais e verticais, e ainda nove janelas. A quantidade de recortes da abordagem Janelas Centralizadas é fixa e igual a nove, sendo que o recorte central representa 50% da área total da região. O objetivo dos recortes retangulares é de capturar características distribuídas diferenciadamente em posições horizontais e verticais da imagem.

A decomposição em diagonais gera sempre quatro recortes contendo cada um dos lados da matriz resultante da divisão pela diagonal principal e secundária. O objetivo é capturar o contraste entre as quatro regiões de canto originais. Esse processo é ilustrado pela Figura 3.9.

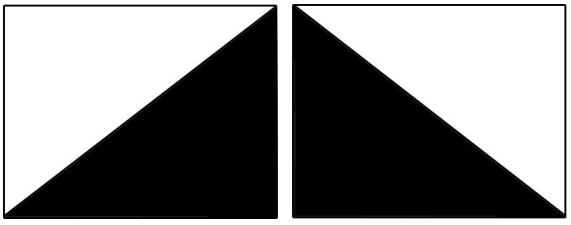


Figura 3.9: Representação da decomposição em diagonais.

A decomposição circular cria novas representações de regiões circulares e concêntricas da região. O objetivo é analisar o incremento da distinção de regiões concêntricas tendo em vista que uma massa possui um núcleo circular mais homogêneo que uma região normal. Um exemplo dessa decomposição é representado na Figura 3.10.

Para calcular o raio de cada círculo, é tomado como base o raio máximo

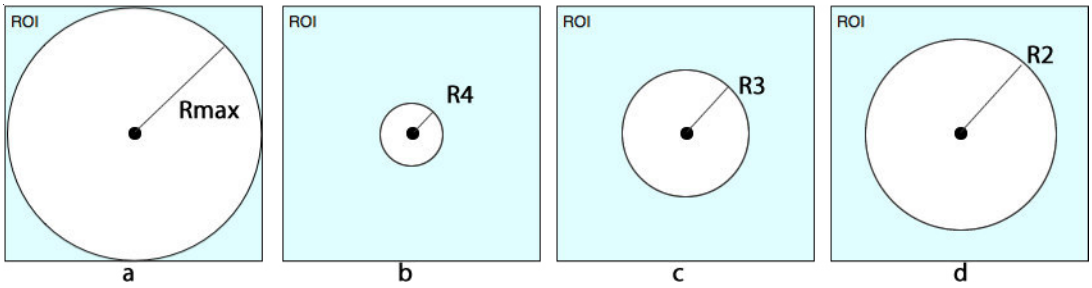


Figura 3.10: Representação da decomposição circular.

determinado pela região, sem englobar áreas externas a ela. Com este fim, é necessário primeiro calcular a menor dimensão da Região de Interesse (ROI). A metade do valor da menor dimensão representa o raio máximo.

A partir do raio máximo, calculam-se os demais através de alteração de proporção em relação ao primeiro, dado pela seguinte razão:

$$R_i = \frac{R_{max}}{i} \tag{3.3}$$

onde $i = 1, 2, 3..., n$ e n é quantidade de raios escolhidos pela aplicação. Para esta metodologia, adotamos $n = 5$. Este parâmetro foi empiricamente testado, onde valores menores normalmente reduzem o desempenho da decomposição e valores maiores não incluem características suficientes para serem adotados.

A decomposição em anéis segue o mesmo principio da decomposição em círculos como exemplificado na Figura 3.11. A diferença básica está no fato de por se tratar de anéis, embora os mesmos compartilham o mesmo centro, eles não possuem intersecção de área. Assim, o raio máximo do anel interno é utilizado como um raio mínimo para o próximo.

A partir do mesmo cálculo do raio máximo obtido pela decomposição de círculos, as Equações 3.4 e 3.5 obtêm os raios mínimos e máximos dos anéis:

$$R_{min}\,(i) = \frac{R_{max}}{n} i \tag{3.4}$$

$$R_{max}\,(i) = \frac{R_{max}}{n}(i+1) \tag{3.5}$$

onde $i = 1, 2, 3...n$ e n é quantidade de raios escolhidos pela aplicação. Para esta metodologia, n tem valor 5 seguindo a mesma definição adotada para a

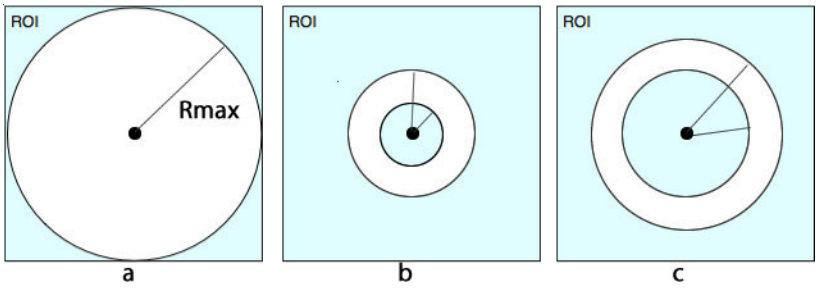


Figura 3.11: Representação da decomposição em anéis.

decomposição em círculos.

Cada decomposição gera recortes individualizados que irão passar por todo o processamento de descrição de textura. Por se tratar de muitos recortes, o seguinte agrupamento de decomposições foi realizado para diminuir a quantidade de bases de características geradas:

- CIRC: contendo as extrações circulares;
- ANEL: contendo as extrações anulares; e
- HVDJ: contendo as extrações resultantes dos processos retangulares e diagonais.

3.4.3 Descrição de Textura

Após as etapas de realce e decomposição, cada região será individualmente descrita usando os índices de diversidade e geoestatística apresentados nas Seções 2.5.2 e 2.5.3, respectivamente:

- Índices de Diversidade: Shannon-Wiener, Simpson, J, Ed, Buzas-Gibson, Camargo, Hill, McIntosh, Diversidade Total, Brillouin, Berger-Parker;
- Índices Geoestatísticos: Moran global e local, Geary global, Getis global, K de Ripley local, *Joint-Count* e *Nearest Neighbor*.

Extração de Características utilizando Índices de Diversidade

Os índices de diversidade são estatísticas que levam em consideração a distribuição das espécies ao longo da região em análise. De uma maneira geral, podemos definir

a distribuição das espécies como sendo o histograma da região sobre os tons de cinza.

Sendo o histograma definido como um vetor onde cada posição representa o quantitativo de ocorrências de um determinado tom de cinza na imagem, então definimos:

- S: é a quantidade de espécies presentes na região é semelhante a quantidade de tons de cinza que tiveram ocorrência maior que zero;
- N: é o número total de pixels da imagem;
- i: é o valor do pixel (espécie), tonalidade; e
- n_i: é o número de pixels com o tom de cinza i.

Assim, através do histograma é possível mapear os itens básicos para a análise de diversidade. Verifica-se em todos os índices de diversidade que sua abrangência é sempre ligada a todos os tons de cinza. De uma maneira geral, tons raros ou com baixa ocorrência perdem representatividade ao serem tratados de maneira conjunta a outros predominantes.

Logo, com intuito de minimizar o efeito da predominância e quantificar os ganhos de representação mais fina da diversidade, propomos nessa metodologia a divisão das espécies em subsistemas de espécies. A hipótese consiste em separá-las e promover maior importância a pequenos grupos populacionais. Essa divisão não é hierárquica. Por fim, é verificada a importância destes subgrupos para a discriminação de massas e tecidos normais.

A divisão é feita de maneira não linear. Verifica-se através do histograma que uma região pode ou não conter todos os tons de cinza e que estes não estão necessariamente agrupados em uma determinada faixa.

Assim, a divisão é realizada conforme o seguinte processo:

- Definir o número de divisões desejadas;
- Obter as espécies;
- Dividir as espécies de acordo com o número de grupos especificados; e
- Realizar o restante da análise em cada grupo separado.

De posse da definição da divisão das espécies, a análise de diversidade da textura é realizada fazendo a combinação dos parâmetros de região espacial e com ou sem divisão de grupos populacionais.

Por fim, cada recorte é submetido a análise de diversidade que compreende aspectos de divisão populacional ou não. Sob cada um, no final, são aplicados os índices de diversidade de Shannon-Wiener, Simpson, J, Ed, Buzas-Gibson, Camargo, Hill, McIntosh, Diversidade Total, Brillouin, Berger-Parker.

Extração de Características utilizando Índices de Geoestatística

Os índices geoestatísticos descrevem a textura em termos de medidas de autocorrelação espacial. A autocorrelação espacial é uma medida de similaridade entre pontos tomados a uma certa distância. Os pontos assumem papéis distintos durante a análise, que correspondem ao papel de referência e analisado. Em todas as abordagens, deve existir ao menos um ponto de referência e em abordagens globais. A maioria dos pontos, em um determinado momento, é tomado como ponto de referência.

Os índices geoestatísticos utilizados para a análise são: Moran local, K de Ripley local, *Joint-Count* e *Nearest Neighbor*.

Esta metodologia trata o aspecto local do índice como sendo o tom de cinza de referência a ser quantificado. Logo, a análise é realizada para um conjunto de pontos de um determinado tom de cinza.

Assumimos que cada tom de cinza possui pontos distribuídos ao longo do espaço da imagem. Utilizamos apenas os que possuem distribuição maior que zero. Todos os demais são desconsiderados.

Se cada tom de cinza é tomado separadamente, então o ponto de referência deve ser escolhido entre os pontos que pertencem a um determinado tom de cinza. Esta metodologia propõe que uma maneira de escolher esse ponto é encontrando o ponto médio da distribuição de pontos. O método escolhido para estimar o ponto médio consiste no processo a seguir, exemplificado na Figura 3.12:

- Estimar o centro de gravidade dos pontos através das médias aritméticas das coordenadas espaciais dos pontos referentes aos eixos de x e y

    1. A percepção de área interna ou externa inexiste quando se analisa

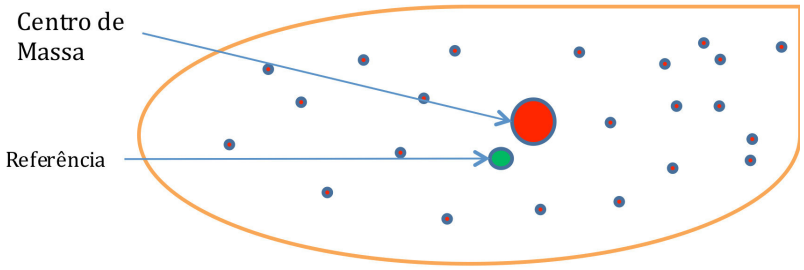


Figura 3.12: Exemplificação do processo de encontrar o ponto médio que será utilizado como referência na extração de características usando geoestatística.

pontos. Assim, o risco que existia ao utilizar centro de gravidade para encontrar o ponto médio e apontar para um ponto fora da região é descartado nessa análise.

2. O ponto médio ou também centro de massa é o representante que minimiza o erro médio quadrático em relação as distâncias dele a todos os outros pontos da análise.

- Com o ponto médio calculado, é necessário materializá-lo na forma de um ponto existente na análise. Para tanto, escolher o ponto presente na análise que esteja mais próximo do ponto médio. A distância é calculada usando distância euclidiana.

Por fim, a quantidade de variáveis geradas para cada análise local é multiplicada pela quantidade de tons presentes na faixa máxima de tons de cinza. Por se tratar de uma imensa quantidade de variáveis, é necessário realizar um processo de seleção de características e também de reorganização das divisões de abordagem.

3.4.4 Descrição Geométrica

Para problemas de descrição de regiões extraídas de mamografias, parte-se do pressuposto que as regiões de massa e normais possuem diferenças de geometria principalmente associadas a forma dos objetos dentro dessas regiões. No caso de regiões de massas, espera-se encontrar formas definidas de círculos ou elipses ambos com contornos não uniformes. No caso de regiões normais espera-se

encontrar formas aleatórias. Partindo dessa hipótese, esta metodologia faz uso de características geométricas para discriminar o padrão presente entre regiões de massas e não massas encontradas no processo de detecção.

Inicialmente investigamos qual o comportamento conjunto de distribuição espacial das regiões em análise. Com esta finalidade, codificamos um mapa de distribuição do conjunto das tonalidades presentes em regiões de massas e normais após a etapa de melhoramento realizada anteriormente. O mapa preserva as relações de localização espacial presentes anteriormente nas imagens, com o intuito de preservar a forma acumulada. Apresentamos o resultado da análise para oito tonalidades distintas através da Figura 3.13. Cada coluna da imagem apresenta a esquerda a distribuição conjunta para amostras normais e a direita para amostras de massa.

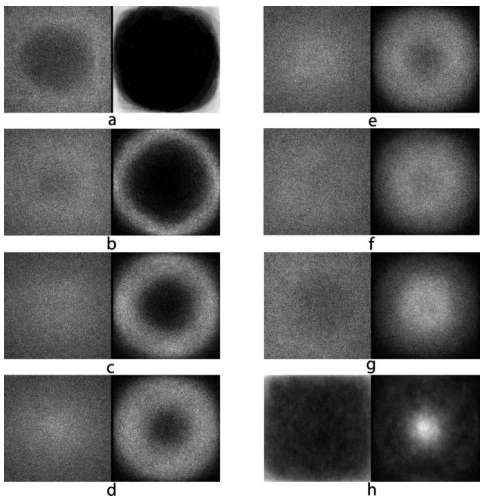


Figura 3.13: Distribuição espacial das tonalidades acumuladas ao longo de todas as amostras normais e de massas para 8 tonalidades chaves, sendo redimensionadas para um tamanho padrão. A esquerda de cada uma das colunas estão regiões normais e à direita regiões de massa. As tonalidades apresentadas são: (a) 32, (b) 64, (c) 96, (d) 128, (e) 160, (f) 192, (g) 224, (f) 255.

Podemos concluir que regiões de massa possuem tons mais altos organizados em seu núcleo. A medida que os tons vão diminuindo de valor, criam-se camadas, com comportamento concêntrico, evidenciando uma característica geométrica das estruturas internas das massas de maneira geral. Já regiões normais apresentam

estruturas dispersas espacialmente que provocam a geração de várias geometrias com uma relação com o centro de baixa evidência circular, evidenciando um comportamento mais aleatório e heterogêneo da textura. Nota-se que existem tons escuros em massas, gerados principalmente pela aplicação da etapa de melhoramento. Esta por sua vez é responsável pela melhor delimitação da forma dos tons mais altos a partir do momento que evita a superconcentração de indivíduos em uma mesma tonalidade.

Por se tratar de uma análise de geometria da distribuição de pontos, a determinação daqueles que serão analisados deve ser tomada como critério de partida. Partimos da ideia de que o conjunto de tonalidades presentes em uma ROI, ordenados por valor é denotado por S. Desejamos subdividir o conjunto S de forma a criar subpopulações $P_1, P_2, ..., P_i...P_n \subset S$ de tonalidade, onde n representa a quantidade de subpopulações.

As seguintes restrições devem ser aplicadas na geração das subpopulações: a) quantidade de elementos em cada subpopulação é igual; b) a interseção entre as subpopulações é nula.

Com esta decomposição podemos quantificar as características espaciais e geométricas dos tons de cinza. Todavia, a informação da tonalidade não é mais necessária. Então, assumimos que a região de interesse (ROI) original é decomposta em N subpopulações de mesmas dimensões, sendo N igual a quantidade de subpopulações. Cada região resultante possui, de maneira binária, a ocorrência espacial de um determinado conjunto de tons de cinza. Um exemplo de decomposição de uma região de massa, apresentada pela Figura 3.14(b), com $N = 3$ é apresentado na Figura 3.14(c-e). A medida que a decomposição agrupa tonalidades mais próximas de 255, maior a concentração circular e centralizada das formas geradas.

A decomposição varia de acordo com a quantidade N escolhida. Como é um parâmetro de ajuste, esta metodologia investiga a utilização do parâmetro N na faixa de 3 a 12.

Cada região decomposta pode ter a geometria determinada usando geometria côncava (Seção 2.5.5). Para tanto, é necessário determinar o parâmetro α. Por haver diferentes tamanhos de regiões, optamos por utilizar a adaptação do parâmetro de acordo com a área da região. Logo, sendo uma janela envolvente da

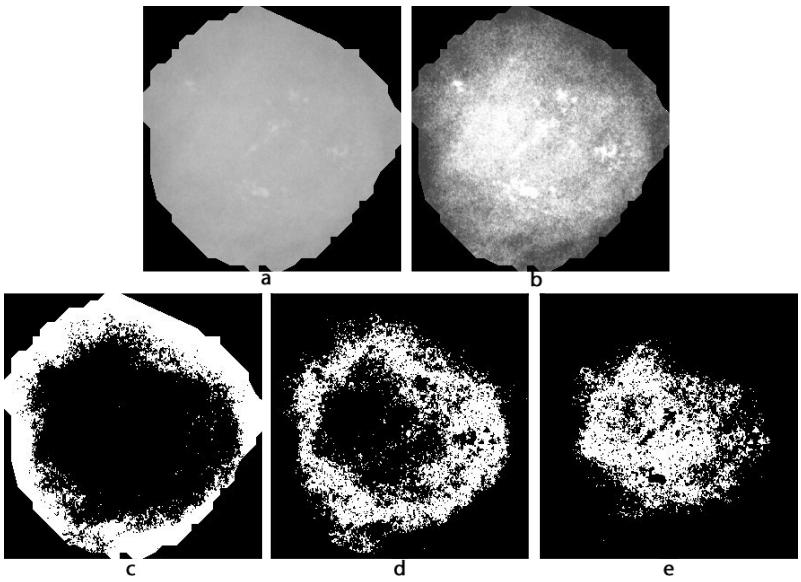


Figura 3.14: Representação da decomposição de uma região de análise em subgrupos espaciais: (a) consiste na imagem original, (b, c, d) resultado binário da primeira, segunda e terceira decomposição, respectivamente.

região de dimensão (X, Y), o valor de α é assumido e calculado por:

$$\alpha = \frac{\sqrt{XY}}{fator} \tag{3.6}$$

onde $fator$ é uma proporção de correção do tamanho máximo de α. Empiricamente, encontramos que um fator que gera bons resultados tem valor 15 que representa a metade do valor médio das diagonais de todas as regiões de interesse. De posse de α, é gerada a geometria de cada uma das regiões decompostas da etapa anterior, correspondendo ao apresentado na Figura 3.15(d-f) para a decomposição apresentada na Figura 3.15(a-c).

Observa-se a formação de vazios internos, bem comuns em geometria côncava e impossíveis de serem determinados se fossem utilizados métodos convexos. Todavia, os buracos representam um problema na interpretação referente ao perímetro do objeto em análise. Para tanto, tratamos qualquer borda como parte do perímetro. Ou seja, os perímetros internos e externos são somados e considerados como somente um.

Se forem formados múltiplos objetos na mesma região, características como perímetro e área são assumidas como a média aritmética das características

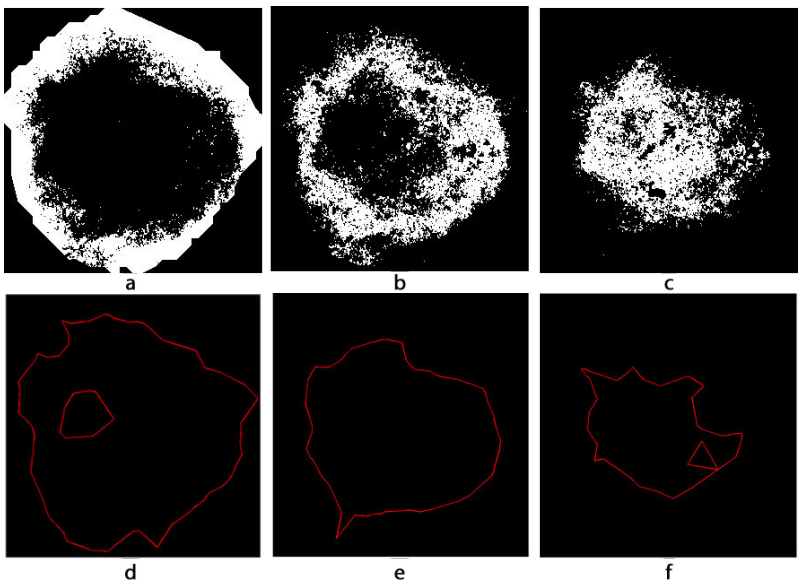


Figura 3.15: Resultado em (d-f) após a aplicação da geometria côncava sobre as regiões apresentadas na Figura 3.14 (c-e) repetidas nessa figura em (a-c).

obtidas individualmente de cada objeto. Assim, cada índice gera somente um valor.

Após a geração das geometrias côncavas de cada subpopulação, utilizamos as medidas apresentadas na Seção 2.5.4 sobre cada subpopulação criada com a decomposição descrita anteriormente.

3.4.5 Reconhecimento

A etapa final da metodologia proposta para redução de falso positivos consiste em classificar cada região em massa ou normal, utilizando reconhecimento de padrões de acordo com a textura ou geometria obtida na etapa de extração de características (Seção 3.4.3 e 3.4.4). Esta metodologia utiliza MVS (Seção 2.6.1) juntamente com o núcleo radial para classificar as regiões.

Uma base é tida como o conjunto de vetores de características resultantes de cada análise gerada anteriormente. O primeiro passo para o reconhecimento é a normalização das variáveis, fundamental para melhorar a convergência do MVS.

Em seguida é realizada a seleção de melhores características para as bases que possuem uma dimensão muito alta. Neste caso, bases geradas a partir de análise

de diversidade e geoestatísticas. A seleção de características é realizada utilizando o algoritmo de ascendência de Hill, através da interface *Greedy Stepwise* (VAFAIE; IMAM, 1994). O algoritmo consiste em uma sequência de passos iterativos, onde o início é dado por uma configuração de solução arbitrária qualquer. Os demais passos trocam uma das variáveis da solução dada na etapa anterior a procura de uma melhor taxa de acerto. Se a troca resultou em sucesso, a nova melhor taxa é anotada. A repetição iterativa acontece até que não ocorra alteração de melhor taxa de acerto.

O terceiro passo consiste na separação de base de treino e teste, realizada por imagem num esquema *Leave-One-Image-Out*. Neste esquema, cada imagem terá suas regiões suspeitas separadas como base de teste pelo menos uma vez, enquanto que todas as outras regiões de todas as outras imagens comporão a base de treinamento. A imagem que é utilizada para teste é alternada, gerando um total de N iterações, onde N é igual ao número de imagens na base.

O quarto passo é para cada uma das bases treino/teste estimar os parâmetros do MVS, C e γ e gerar um modelo de aprendizado o qual no quarto passo de predição será usado para classificar as regiões descritas na base de teste.

Um outro parâmetro utilizado é o grau regulador da penalidade por erro em uma determinada classe. Esse parâmetro corrige distorções de distribuição de amostras. Em todos os modelos gerados, esse parâmetro tem o valor igual a 5, que representa aproximadamente o grau de desbalanceamento entre massas e não massas.

Para avaliar os resultados do reconhecimento e também da metodologia como um todo, são utilizados as medidas de Taxa Geral de Acerto de Verdadeiros Positivos por Imagem (S_{img}), Taxa Média de Falso Positivos por Imagem (Fp/i). Para avaliar a capacidade de generalização do modelo criado pelo classificador é utilizado o Número de Vetores de Suporte do modelo (quanto maior for este valor, mais próximo de *overfit* está o modelo gerado) e a curva FROC.

Capítulo 4

Resultados e Discussões

Este capítulo apresenta os resultados obtidos usando as bases públicas de mamografias MIAS e DDSM. O resultado de cada etapa da metodologia é revisado individualmente e discutido de maneira a fomentar as conclusões do capítulo seguinte e também as melhorias que a metodologia pode absorver.

4.1 Resultados usando MIAS

Os testes realizados nessa seção envolvem a utilização das 74 mamografias da base MIAS. Ao todo foram selecionadas todas as mamografias que tinham ao menos uma lesão de massa informada. As imagens de mamografia obtidas diferem em classe de densidade observada pelo especialista e são classificadas pela base em 3 categorias: *Fatty* ou pouco densa; *Granular* ou de densidade média e *Dense* para representar alta densidade. A distribuição das imagens pelas categorias é dada por:

- Fatty: 29
- Granular: 26
- Dense: 19

Inicialmente, todas foram submetidas aos processos de realce, explicados na Seção 3.2, que possuem parâmetros ajustados pelo contraste. Na etapa de detecção, inicialmente é aplicado o *MeanShift* e filtro STD sobre as imagens,

conforme Seção 3.3 descreve. Os passos seguintes são o agrupamento incremental para ajuste dos parâmetros do FSA, detecção usando FSA e redução de falso positivos usando índices de diversidade, geoestatísticos e geometria côncava.

Para a etapa de detecção de regiões suspeitas, inicialmente é realizado o ajuste dos parâmetros do algoritmo FSA usando a clusterização incremental das imagens, baseado na informação de contraste.

O agrupamento incremental necessita de um parâmetro de otimização que é a distância máxima intra grupo. Inicialmente executamos uma série de testes para melhor entender o funcionamento do parâmetro. Em todos os testes, a taxa de sensibilidade se manteve constante, em 97,30%. Assim, o critério de otimização passa a ser o número médio de falso positivos e quantidade de grupos gerados. Ambos devem ser reduzidos. Ao final, a configuração individual que atende a estes requisitos utiliza distância igual a 15, com um total de 17 grupos e 4,094 falso positivos por imagem.

Após a escolha dos parâmetros, o FSA é aplicado de maneira a identificar regiões suspeitas que consistem em regiões de alta densidade de tonalidades claras. Para as 74 imagens da base MIAS utilizadas, a quantidade de regiões geradas pela etapa de detecção foi de 84 massas e 303 não massas, totalizando 391 regiões. Em 72 das 74 imagens foi identificado ao menos uma região que tenha intersecção com a região anotada pelo especialista como massa, correspondendo a uma sensibilidade de 97,29%.

Independente da configuração escolhida na etapa de agrupamento incremental, duas imagens não apresentam uma região suspeita que seja a massa. Essas massas foram perdidas nas etapas anteriores da metodologia e são analisadas mais a frente no estudo de casos. Mesmo não tendo apresentado regiões de massa, outras regiões que são falso positivos podem ter sido geradas para estas imagens e portanto continuam sendo consideradas em todas as etapas da metodologia.

4.1.1 Primeira Redução de Falso Positivos

Com as regiões suspeitas segmentadas, a próxima etapa da metodologia consiste na redução de falso positivos usando análise de diversidade, geoestatísticas e geometria côncava. Cada análise de redução de falso positivos é realizada separada por índice/abordagem utilizada, sempre sobre a mesma base de regiões e seguindo

a metodologia apresentada na Seção 3.4.

As Tabelas 4.1, 4.2, 4.3, 4.4, 4.5 apresentam os resultados ordenados por maior sensibilidade por Imagem (S_{img}), menor média de falso positivos por imagem (Fp/i) e menor quantidade de vetores de suporte (nSv). A acurácia total (Acc) do classificador também é apresentada. Esta se refere a acurácia de classificação das regiões submetidas ao classificador e não aos exames propriamente dito. Nas abordagens que utilizam índices de diversidade ou geoestatística, os resultados estão classificados por abordagem de decomposição espacial, conforme Seção 3.4.2 e quanto a existência ou não de subdivisão de população.

Todas as abordagens mantiveram a mesma taxa de sensibilidade anterior à etapa de redução de falso positivo. De uma forma geral, o melhor resultado para redução de falso positivo é encontrado usando geometria côncava, quando é utilizada a divisão em 3 subpopulações, obtendo 100% de acerto em verdadeiros positivos e de redução de falso positivo. Todos os demais resultados usando geometria côncava também são melhores do que qualquer resultado obtido usando análise de diversidade e geoestatística.

O melhor resultado geral usando índices de diversidade foi encontrado utilizando o índice diversidade total na abordagem em ANEL com média de 0,77 falso positivos gerados por imagem. Quando utilizando índices geoestatísticos, o melhor resultado foi obtido usando a estatística *Joint-Count*, na abordagem ANEL, com 0,59 em média de falso positivos por imagem. Os dois resultados, embora significantes, possuem um problema relacionado com o grau de generalização do classificador.

Como a abordagem utilizada para geração do modelo de treinamento foi o *Leave-One-Image-Out*, é fundamental que a quantidade média de vetores de suporte presentes no treinamento (nSv) permaneça baixa representando a qualidade do modelo gerado. Quanto maior esse índice, mais superajustado se encontra o modelo treinado e, por conseguinte, os resultados não são confiáveis. O índice é uma proporção em relação ao tamanho da base de treinamento.

No caso da utilização da geometria côncava, o melhor resultado tem nSv = 23,82. Considerando uma base de treinamento com 380 indivíduos em média, o modelo gerado pela geometria côncava necessitou de apenas 6,3% dos vetores de treinamento para geração do aprendizado. Concluímos que este é um modelo de

alto desempenho e generalista. Todavia, quando verificamos o mesmo índice sobre os testes realizados com índice de diversidade ou geoestatística, observamos um alto grau de associação entre a base de treinamento e os vetores que se encontram no modelo.

O mesmo acontece a medida que o número de subpopulações aumenta para a geometria côncava. Concluímos que quanto maior a quantidade de características maior a dificuldade para o MVS gerar um classificador de baixa complexidade, quando utilizando medidas obtidas pela abordagem em geometria côncava.

4.1.2 Segunda Redução de Falso Positivos

A segunda redução de falso positivos consiste em realimentar a metodologia com as regiões detectadas como massas na primeira redução de falso positivos. Caso a região possua mais do que uma estrutura interna, essa será subdividida para gerar objetos individuais.

Apresentamos aqui os resultados obtidos usando como base de entrada a saída da base gerada pela Geometria Côncava. Sobre esta é realizado a descrição por Geometria Côncava por se tratar do melhor resultado obtido na primeira etapa. Cada base representa o resultado oriundo de cada teste de subpopulação gerado na primeira etapa de redução de falso positivos. Onde existiam múltiplos objetos numa mesma região agrupados, foi realizado a divisão, gerando um novo número de falso positivos médio por imagem como representado na Tabela 4.6.

Por se tratar de um grande número de combinações (10 subpopulações bases por 10 subpopulações da nova análise resultando em 100 novas análises), a Tabela 4.7 apresenta apenas os 10 melhores e 10 piores resultados para as Bases 3-12, com as subpopulações variando de 3 a 12 subpopulações.

Verificamos que a segunda redução de falso positivos reduz ao menos um terço a quantidade de falso positivos, principalmente usando novamente geometria côncava com três divisões sub populacionais, mantendo inalterado a sensibilidade da metodologia. Os novos índices médios de falso positivos para as abordagens geométricas se situam na faixa de 0,33 à 0,944. Assim, houve menos que 1 falso positivo por imagem em qualquer situação.

Tabela 4.1: Resultados obtidos para abordagem de decomposição ANEL e índices de diversidade para a base MIAS.

Índice	Divisão	S_{img} (%)	FP/i	Acc (%)	nSv
Diversidade Total	sem	97,30	0,77	84,88	294
Buzas Gibson	com	97,30	0,88	82,76	272
Buzas Gibson	sem	97,30	0,89	82,49	260
McIntosh	sem	97,30	0,91	82,23	253
J	sem	97,30	0,92	81,96	303
Diversidade Total	com	97,30	1,00	80,37	278
Brillouin	sem	97,30	1,03	79,84	281
Camargo	sem	97,30	1,11	78,25	279
Hill	sem	97,30	1,11	78,25	290
ED	sem	97,30	1,12	77,98	269
Shannon	sem	97,30	1,14	77,72	251
Simpson	sem	97,30	1,19	76,66	262
Brillouin	com	97,30	1,20	76,39	279
Berger Parker	sem	97,30	1,23	75,86	267
Berger Parker	com	97,30	1,31	74,27	291
J	com	97,30	1,42	72,15	279
Hill	com	97,30	1,43	71,88	290
Camargo	com	97,30	1,43	71,88	295
ED	com	97,30	1,45	71,62	288
McIntosh	com	97,30	1,57	69,23	289
Shannon	com	97,30	1,58	68,97	273
Simpson	com	97,30	1,61	68,44	301

4.2 Resultados usando DDSM

Os testes realizados nessa seção envolvem a utilização de 621 mamografias da base DDSM selecionadas aleatoriamente dos volumes benigno 01-05 e câncer 01-05. A distribuição das imagens pelas categorias de densidade BI-RADS é dada por:

- BI-RADS 1: 121

Tabela 4.2: Resultados obtidos para abordagem de decomposição CIRC e índices de diversidade para a base MIAS.

Índice	Divisão	S_{img} (%)	FP/i	Acc (%)	nSv
J	sem	97,30	0,8514	83,29	279
Buzas Gibson	sem	97,30	0,8649	83,02	250
Diversidade Total	sem	97,30	0,8784	82,76	258
Buzas Gibson	com	97,30	0,9595	81,12	276
McIntosh	sem	97,30	1	80,37	255
Diversidade Total	com	97,30	1	80,32	286
Berger Parker	com	97,30	1,025	79,70	267
Brillouin	com	97,30	1,0405	79,52	263
Brillouin	sem	97,30	1,0541	79,31	269
Camargo	com	97,30	1,0676	78,99	248
ED	sem	97,30	1,0946	78,51	269
Simpson	sem	97,30	1,1081	78,25	270
Camargo	sem	97,30	1,1216	77,98	277
McIntosh	com	97,30	1,1486	77,39	264
Hill	sem	97,30	1,1757	76,92	277
Shannon	sem	97,30	1,2162	76,13	255
Berger Parker	sem	97,30	1,2297	75,86	273
Hill	com	97,30	1,2297	75,80	280
J	com	97,30	1,2703	75,00	265
ED	com	97,30	1,2838	74,73	279
Shannon	com	97,30	1,3108	74,20	274
Simpson	com	97,30	1,3108	74,20	306

- BI-RADS 2: 267
- BI-RADS 3: 171
- BI-RADS 4: 62

Após a etapa de pré-processamento, as imagens foram submetidas a etapa de detecção. Seguindo os critérios utilizados pelo agrupamento incremental para

Tabela 4.3: Resultados obtidos para abordagem de decomposição HVDJ e índices de diversidade para a base MIAS.

Índice	Divisão	S_{img} (%)	FP/i	Acc (%)	nSv
Berger Parker	com	97,30	0,7703	84,84	275
J	sem	97,30	0,8784	82,71	262
Buzas Gibson	com	97,30	0,9189	81,91	273
Hill	sem	97,30	0,9189	81,91	282
Diversidade Total	com	97,30	0,973	80,85	295
McIntosh	com	97,30	1,0135	80,05	240
Shannon	sem	97,30	1,0135	80,05	265
Simpson	com	97,30	1,0135	80,05	284
Brillouin	com	97,30	1,027	79,79	271
Buzas Gibson	sem	97,30	1,0676	78,99	279
Simpson	sem	97,30	1,0946	78,46	271
Diversidade Total	sem	97,30	1,0946	78,46	304
Hill	com	97,30	1,1081	78,19	274
Camargo	com	97,30	1,1216	77,93	289
J	com	97,30	1,1486	77,39	289
Brillouin	sem	97,30	1,1622	77,13	285
ED	sem	97,30	1,1622	77,13	305
Shannon	com	97,30	1,2432	75,53	277
Ed	com	97,30	1,2432	75,53	309
Berger Parker	sem	97,30	1,2703	75,00	276
McIntosh	sem	97,30	1,2973	74,47	281
Camargo	sem	97,30	1,3243	73,94	301

seleção de parâmetros, a distância utilizada para agrupamento foi de 49, a qual gerou sensibilidade de 91,63% (a detecção falhou em 52 imagens, detalhadas por grupo de densidade na Tabela 4.8) com taxa média de falso positivos de 3,76 e 12 grupos.

Para as 621 imagens da base DDSM utilizadas, a quantidade de regiões geradas pela etapa de detecção foi de 957 massas e 2338 não massas, totalizando 3295 regiões. O resultado inferior, em termos de sensibilidade se comparado com o

Tabela 4.4: Resultados obtidos usando índices geoestatísticos sobre as abordagens ANEL, CIRC e HVDJ para a base MIAS.

Índice	Abordagem	$S_{img}(\%)$	FP/i	Acc (%)	nSv
Joint Count	ANEL	97,30	0,5946	88,30	261
NNA	HVDJ	97,30	0,9324	81,65	310
Local Ripley	ANEL	97,30	0,9459	81,38	296
Local Ripley	HVDJ	97,30	0,9595	81,12	282,91
Joint Count	HVDJ	97,30	0,973	80,85	267
Local Ripley	CIRC	97,30	1,027	79,79	291
NNA	ANEL	97,30	1,1216	77,93	285
NNA	CIRC	97,30	1,1757	76,86	289
Joint Count	CIRC	97,30	1,3919	72,61	262
Local Moran	CIRC	97,30	1,6757	67,02	331
Local Moran	ANEL	97,30	1,7297	65,96	324
Local Moran	HVDJ	97,30	1,7568	65,43	301

Tabela 4.5: Resultados obtidos para redução de falso positivos utilizando geometria côncava para a base MIAS.

Subpopulações	S_{img} (%)	FP/i	Acc (%)	nSv
3	97,30	0	100,00	23
4	97,30	0,0541	98,94	56
8	97,30	0,0541	98,94	81
9	97,30	0,1081	97,87	89
5	97,30	0,1081	97,87	97
6	97,30	0,1351	97,34	77
10	97,30	0,1757	96,54	91
7	97,30	0,1757	96,54	120
11	97,30	0,2568	94,95	135
12	97,30	0,3243	93,62	122

testes realizados com o MIAS, indica o maior grau de dificuldade na detecção de massas no DDSM haja visto o grande aspecto de variância em termos de localização, definição e clareza da lesão. Os principais motivos de erro da

Tabela 4.6: Novas bases geradas após a aplicação da primeira redução de falso positivos, usando geometria côncava, para a base MIAS.

Base	Fp/i Primeira Etapa	Após divisão		
		Massas	Não massas	FP/i
3	0	93	99	1,337
4	0,0541	93	106	1,432
5	0,0541	93	111	1,500
6	0,1081	93	113	1,527
7	0,1081	93	128	1,729
8	0,1351	93	103	1,391
9	0,1757	93	118	1,594
10	0,1757	93	125	1,689
11	0,2568	93	130	1,756
12	0,3243	93	147	1,986

metodologia para detecção de massas foram massas de tamanho muito reduzido e baixa intensidade em relação a outros tecidos da mama. Estas são duas restrições impostas pela metodologia nas extensões do FSA e que devem ser motivos de melhoria em trabalhos futuros.

4.2.1 Primeira Redução de Falso Positivos

As Tabelas 4.9, 4.10, 4.11, 4.12, 4.13 apresentam os resultados ordenados por maior sensibilidade por Imagem (S_{img}), menor média de falso positivos (Fp/i) e menor quantidade de vetores de suporte (nSv). A acurácia total (Acc) do classificador também é apresentada. Esta se refere a acurácia de classificação das regiões submetidas ao classificador e não aos exames propriamente dito. Nas abordagens que utilizam índices de diversidade ou geoestatística, os resultados estão classificados por abordagem de decomposição espacial, conforme Seção 3.4.2 e quanto a existência ou não de subdivisão de população.

O melhor resultado geral foi obtido usando geometria côncava, obtendo sensibilidade de 91,63%, classificando corretamente todos as regiões normais (número nulo de falso positivos) com a menor quantidade de vetores de suporte

Tabela 4.7: Resultados obtidos na segunda redução de falso positivos usando geometria côncava para a base MIAS.

Ordem	Base	Subpopulações	S_{img} (%)	FP/i	Acc (%)	nSv
1	8	3	97,30	0,3333	87,76	94
2	6	3	97,30	0,3333	88,35	101
3	4	3	97,30	0,3472	87,44	98
4	5	3	97,30	0,3472	87,75	100
5	3	3	97,30	0,3611	86,46	90
6	11	3	97,30	0,4306	86,10	108
7	9	3	97,30	0,4722	83,89	94
8	4	4	97,30	0,4722	82,91	109
9	6	4	97,30	0,4722	83,50	111
10	5	4	97,30	0,4722	83,33	115
91	9	7	97,30	0,8194	72,04	167
92	10	7	97,30	0,8333	72,48	160
93	11	7	97,30	0,8333	73,09	168
94	7	11	97,30	0,8333	72,85	169
95	12	9	97,30	0,8611	74,17	174
96	12	4	97,30	0,8889	73,33	128
97	7	12	97,30	0,9167	70,14	163
98	12	7	97,30	0,9167	72,50	170
99	7	7	97,30	0,9306	69,68	175
100	12	8	97,30	0,9444	71,67	181

Tabela 4.8: Relação de erros por classe de densidade para os testes realizados no DDSM

Densidade	Total Erros (Qtd.)	Perc. intra densidade	Perc. extra densidade
1	6	4,96%	11,54%
2	24	8,99%	46,15%
3	16	9,36%	30,77%
4	6	9,68%	11,54%

observado para todas as abordagens (1731) logo o classificador mais robusto. Todavia, ao contrário do que aconteceu quando testado com a base MIAS, todas as abordagens tiverem seu desempenho melhorado em termos de redução de falso positivo. Outros experimentos usando índices de diversidade obtiveram taxa máxima de sensibilidade e de redução de falso positivos, tornando como critério de desempate o número de vetores de suporte.

Tabela 4.9: Resultados obtidos para abordagem de decomposição ANEL e índices de diversidade para a base DDSM.

Índice	Divisão	S_{img} (%)	FP/i	Acc (%)	nSv
Brillouin	com	91,63	0,000	100,00	2313
Berger Parker	com	91,63	0,002	99,97	3224
ED	com	91,63	0,005	99,91	2226
Buzas Gibson	com	91,63	0,071	98,66	2705
Shannon	com	91,63	0,090	98,24	2570
Diversidade Total	sem	91,63	0,346	93,47	3204
Hill	com	91,63	0,486	90,71	2216
J	com	91,63	0,651	87,68	3005
Shannon	sem	91,47	0,871	83,49	2499
Diversidade Total	com	91,47	1,483	71,90	2932
J	sem	91,47	3,021	43,03	3165
Camargo	com	91,30	0,501	90,44	2326
Simpson	com	91,30	0,762	85,58	2558
Simpson	sem	91,30	2,939	44,52	3052
Buzas Gibson	sem	91,14	0,035	99,21	2682
Camargo	sem	91,14	3,361	36,42	2868
ED	sem	90,98	1,905	63,55	2441
Berger Parker	sem	90,82	2,686	49,20	3173
McIntosh	com	90,66	1,688	67,65	2520
McIntosh	sem	90,66	2,789	47,10	2673
Hill	sem	90,50	1,733	67,04	2318
Brillouin	sem	90,02	2,916	44,13	2832

Tabela 4.10: Resultados obtidos para abordagem de decomposição CIRC e índices de diversidade para a base DDSM.

Índice	Divisão	S_{img} (%)	FP/i	Acc (%)	nSv
Berger Parker	com	91,63	0,002	99,97	3187
Brillouin	com	91,63	0,005	99,91	2050
Diversidade Total	sem	91,63	0,019	99,64	2838
Shannon	com	91,63	0,069	98,69	2658
Buzas Gibson	com	91,63	0,396	92,50	2874
ED	sem	91,63	0,884	83,28	2673
ED	com	91,63	1,448	72,38	2914
Hill	sem	91,47	0,520	90,14	1879
Diversidade Total	com	91,47	1,145	78,21	2064
Simpson	com	91,30	0,596	88,71	3099
Shannon	sem	91,14	0,750	85,55	1880
Camargo	com	91,14	0,857	83,58	2230
Camargo	sem	91,14	1,176	77,42	2186
McIntosh	sem	91,14	2,776	47,56	3188
Berger Parker	sem	90,98	2,675	49,44	3192
Simpson	sem	90,98	3,014	42,94	3058
Buzas Gibson	sem	90,66	0,338	93,17	2018
Brillouin	sem	90,66	1,684	67,71	2449
Hill	com	90,50	1,309	74,69	2392
J	com	90,50	1,636	68,29	2528
McIntosh	com	90,50	1,900	63,67	2888
J	sem	89,37	2,454	52,75	2576

Comparando com os experimentos realizados com o MIAS verifica-se os seguintes comportamentos:

1. Redução da sensibilidade: explicada pela grande variedade de imagens do DDSM; e

2. Melhoria da qualidade da redução de falso positivos: a grande variedade também introduziu mais indivíduos no classificador, disponibilizando mais

Tabela 4.11: Resultados obtidos para abordagem de decomposição HVDJ e índices de diversidade para a base DDSM.

Índice	Divisão	S_{img} (%)	FP/i	Acc (%)	nSv
Diversidade Total	sem	91,63	0,000	100,00	2730
Berger Parger	com	91,63	0,000	100,00	2569
ED	sem	91,63	0,018	99,67	2268
Diversidade Total	com	91,63	0,019	99,64	1946
Simpson	sem	91,63	0,372	92,98	1958
Buzas Gibson	sem	91,63	2,163	58,99	2673
Brillouin	com	91,63	2,380	55,10	3174
ED	com	91,63	2,831	46,60	3004
McIntosh	com	91,63	3,045	42,56	3275
McIntosh	sem	91,63	3,101	41,49	3277
Brillouin	sem	91,63	3,240	38,88	3288
Camargo	com	91,63	3,451	34,81	2849
J	com	91,63	3,651	31,08	2856
Camargo	sem	91,63	3,738	29,50	2837
J	sem	91,63	3,738	29,50	2825
Shannon	com	91,47	0,140	97,27	2020
Buzas Gibson	com	91,47	2,238	57,75	2736
Shannon	sem	91,47	3,623	31,53	2835
Berger Parger	sem	91,30	1,858	64,79	2495
Hill	sem	90,82	3,118	40,89	2771
Hill	com	90,66	2,531	51,79	2994
Simpson	com	90,66	3,077	41,56	2872

suporte para um aprendizado eficiente.

O primeiro item ficou evidente quando a etapa de redução de falso positivos não manteve constante a taxa de sensibilidade, mesmo que a classe de massa tenha sido priorizada, conforme o mesmo experimento realizado no MIAS. A maior variedade de formas, tamanho, qualidade dos exames provocou essa alteração que é comum em metodologias testadas com a base DDSM.

Tabela 4.12: Resultados obtidos usando índices geoestatísticos sobre as abordagens ANEL, CIRC e HVDJ para a base DDSM.

Índice	Abordagem	S_{img} (%)	FP/i	Acc (%)	nSv
Joint Count	CIRC	91,63	0,250	95,24	1954
Joint Count	HVDJ	91,63	1,535	70,74	2494
Local Moran	ANEL	91,63	3,011	43,25	2670
Local Moran	CIRC	91,63	3,396	35,99	2770
Local Ripley	CIRC	91,47	0,256	95,14	1995
Local Moran	HVDJ	90,82	0,823	84,26	3089
NNA	CIRC	90,34	0,224	95,11	2457
Joint Count	ANEL	89,86	0,137	96,48	2094
NNA	ANEL	89,86	2,931	44,04	3217
Local Ripley	HVDJ	88,73	0,190	95,23	2208
Local Ripley	ANEL	85,83	0,398	89,32	2557

Tabela 4.13: Resultados obtidos para redução de falso positivos utilizando geometria côncava para a base DDSM.

Subpopulações	S_{img} (%)	FP/i	Acc (%)	nSv
4	91,63	0,000	100,00	1731
3	91,63	0,000	100,00	2163
6	91,63	0,021	99,61	2185
5	91,47	0,016	99,61	2010
12	90,98	0,857	83,00	2495
7	89,37	1,219	75,84	2031
11	89,37	1,235	75,27	2106
8	89,21	1,399	72,17	2067
10	88,08	1,667	65,83	2259
9	87,12	1,486	69,20	2137

Embora a variedade provoque um aumento da dificuldade da classificação de massas, o segundo item analisa que esta provocou um melhor treinamento das regiões que não são massas e melhoria da redução de falso positivos.

4.2.2 Segunda Redução de Falso Positivos

Conforme realizado nos testes com o MIAS, na Tabela 4.14 apresentamos os novos valores de distribuição de massas e não massas após a redução da primeira etapa para os melhores resultados obtidos na primeira etapa que são aqueles que usaram Geometria Côncava e obtiveram taxa máxima de sensibilidade. Uma observação importante da tabela é que a taxa de falso positivos não cresceu significativamente indicando que houve uma pequena junção de regiões normais com massas.

Tabela 4.14: Novas bases geradas após a aplicação da primeira redução de falso positivos, usando geometria côncava, para a base DDSM.

Base	Fp/i Primeira Etapa	Após divisão		
		Massas	Não massas	FP/i
3	0	1026	61	0,0982
4	0	1026	61	0,0982
6	0,021	1026	74	0,1191

A Tabela 4.15 apresenta apenas os resultados para a segunda redução de falso positivos para as Bases 3, 4 e 6, com as subpopulações variando de 3-12. Os resultados estão ordenados por maior sensibilidade, menor Fp/i e menor nSV. Todos os testes atingiram taxa máxima de sensibilidade. Os dois primeiros resultados, que usaram as subpopulações 7 e 6 da base 3 obtiveram 100% de redução de falso positivos. No entanto, subentende-se que estes não sejam os melhores resultados já que a quantidade de vetores de suporte (nSv) é muito alta.

Portanto, o melhor resultado é encontrado usando a base 4, com decomposição 11 atingindo uma taxa média de falso positivos de 0,013, com uma quantidade de vetores de suporte por volta de 30%. Neste caso, além de reduzir significantemente os falso positivos ao mesmo tempo que mantém a sensibilidade, também produz um classificador robusto.

4.3 Estudo de Casos

Com intuito de melhor interpretar os resultados encontrados, esta seção apresenta o estudo individual de alguns casos para exemplificar os experimentos como um

todo. Para tanto, utilizaremos como referência os resultados obtidos pela melhor configuração de teste da nossa metodologia para a base de dados específica.

Casos de Sucesso

Inicialmente vamos interpretar os resultados encontrados para a imagem mdb005 da base MIAS, caso modelo apresentado durante o Capítulo 3 para exemplificação do processo. Apresentamos na Figura 4.1, todas as etapas da metodologia aplicada sobre a imagem original (a), sendo em (b) o resultado obtido pelo *MeanShift*, sobre a imagem pré-processada, em (c) o resultado obtido após o filtro STD, em (d) o resultado obtido após o FSA, em (e) após a primeira redução de falso positivos e em (f) após a segunda redução de falso positivos.

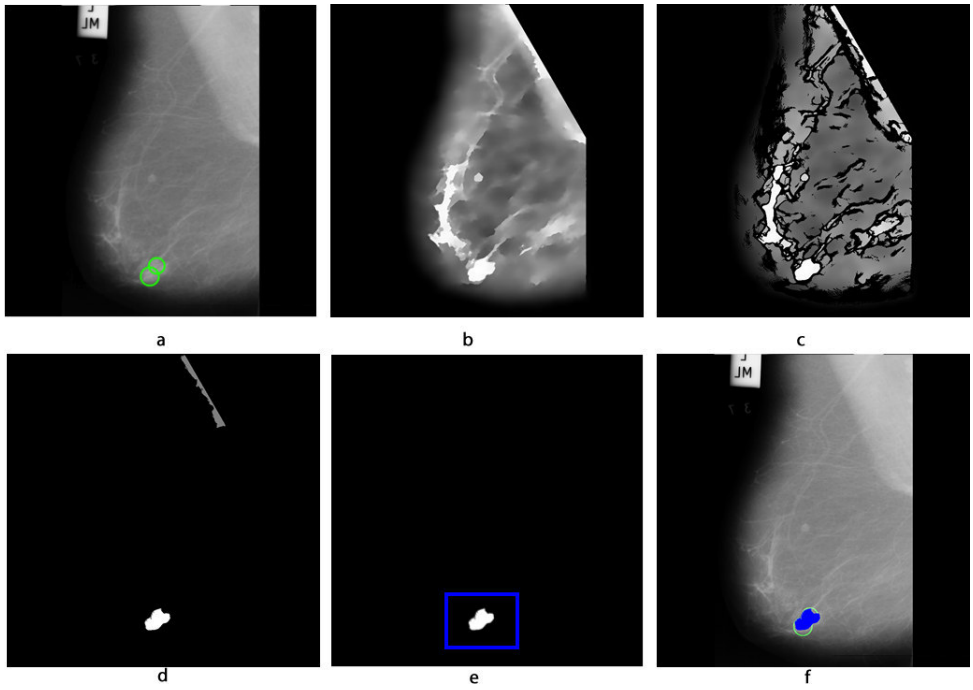


Figura 4.1: Estudo de caso para a imagem mdb005: (a) imagem original, (b) após a etapa de melhoramento, (c) após *MeanShift* e Filtro STD, (d) após o *Fast Scanning*, (e) após a primeira redução de falso positivos e em (f) após a segunda redução de falso positivos.

Este primeiro caso é uma imagem de densidade *Fatty* (ou mais baixa densidade), onde claramente a região da massa se destaca em relação ao tecido

da glândula mamária. Por esse motivo, também possui alto contraste calculado (equivalente a 417,56). Existem duas marcações de massas assinaladas, inclusive com uma pequena sobreposição. O resultado final da detecção gerou 1 massa e 1 região normal apresentado pela Figura 4.1(d).

As duas regiões suspeitas presentes são então submetidas ao processo a primeira redução de falso positivos, que elimina a região normal próxima ao músculo peitoral e mantém a região de massa, obtendo 100% de acerto. A segunda redução de falso positivos mantém a massa, conforme resultado final apresentado pela Figura 4.1(f).

As Figuras 4.2 e 4.3 apresentam casos de sucesso para a base DDSM. Em ambas, houve uma pequena quantidade de falso positivos gerada. A Figura 4.2 apresenta um exemplo onde a região de massa não foi a mais densa da imagem (analise a cor mais cinza em Figura 4.2(b)). Nesse caso, a extensão por pertinência em relação a região mais densa foi preponderante para a inclusão da massa no conjunto de regiões suspeitas.

Caso de grande redução de falso positivos

Neste caso de estudo, analisamos o potencial de redução de falso positivos. Algumas imagens estão em grupos com uma quantidade maior de imagens e portanto parâmetros de detecção menos precisos. Normalmente são imagens de contraste mediano. Exemplos deste tipo de imagem são apresentados através da: Figura 4.4 e Figura 4.5.

Verificamos a presença de um grande número de falso positivos (Figura 4.4(b)), que poderia ser muito maior caso não houvesse a união como uma etapa do FSA. Foram detectados como regiões suspeitas o total de 1 massa e 9 normais, bem superior a média apresentada pelo detector. Após a etapa de redução de falso positivos, resta apenas a região de massa, apresentada em Figura 4.4(c). Todavia, agregada a massa, existem outras duas estruturas, que por semelhança e proximidade em relação a massa, acabaram por serem tratadas como somente uma.

Embora exista a agregação de tecidos que não são massa em regiões de massa, podemos analisar a vantagem da etapa de junção no sentido da união de informações de textura suficientemente discriminatórias para produzir a redução

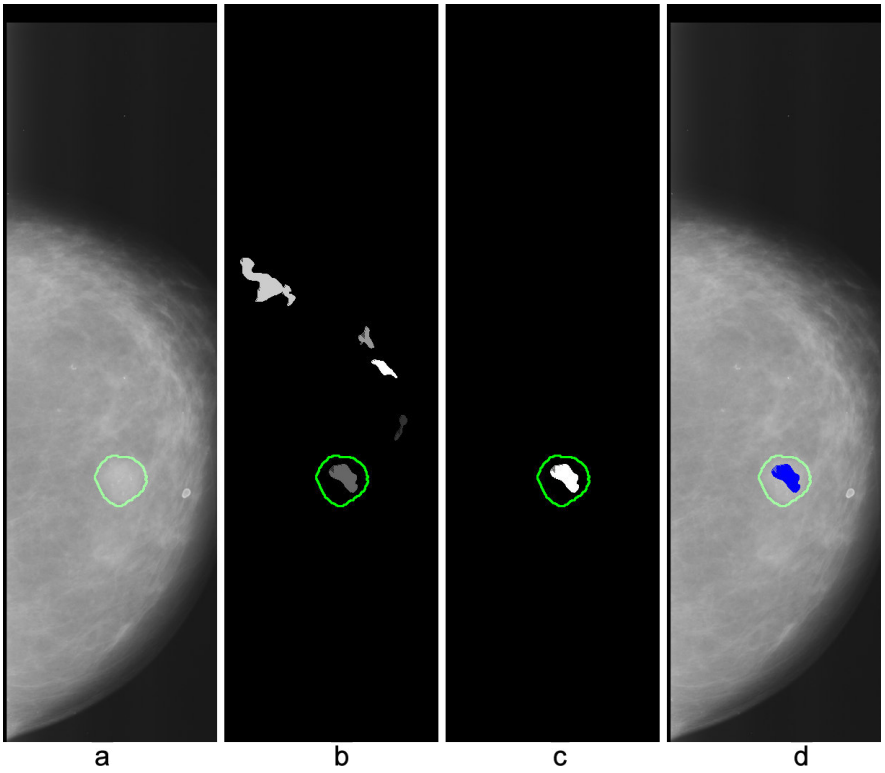


Figura 4.2: Estudo de caso para a imagem B3091 LEFT CC: (a) imagem original, (b) após a detecção de regiões suspeitas, (c) resultado da primeira redução de falso positivos, (d) resultado final após a segunda redução de falso positivos.

de falso positivos.

Com a quantidade de informação textural realçada, métodos como os descritos na geometria côncava se valem de mecanismos mais eficientes de representação conjunta de forma e textura. Se não houvesse união, analisando apenas forma, existiriam na Figura 4.4(b) muitas estruturas semelhantes a círculos além da massa. Ainda, qualquer estrutura no centro da mamografia tem praticamente a mesma densidade e intensidade da massa, por se tratar de uma imagem de densidade granular.

A Figura 4.5 apresenta um caso de grande redução de falso positivos e também de super segmentação da massa. A baixa densidade da região de massa fez com que a maior parte de sua área fosse eliminada. Por este motivo, o ajuste de parâmetros para esta imagem permitiu a inclusão de uma grande quantidade de regiões de não massa, eliminadas durante a etapa de redução falso positivos.

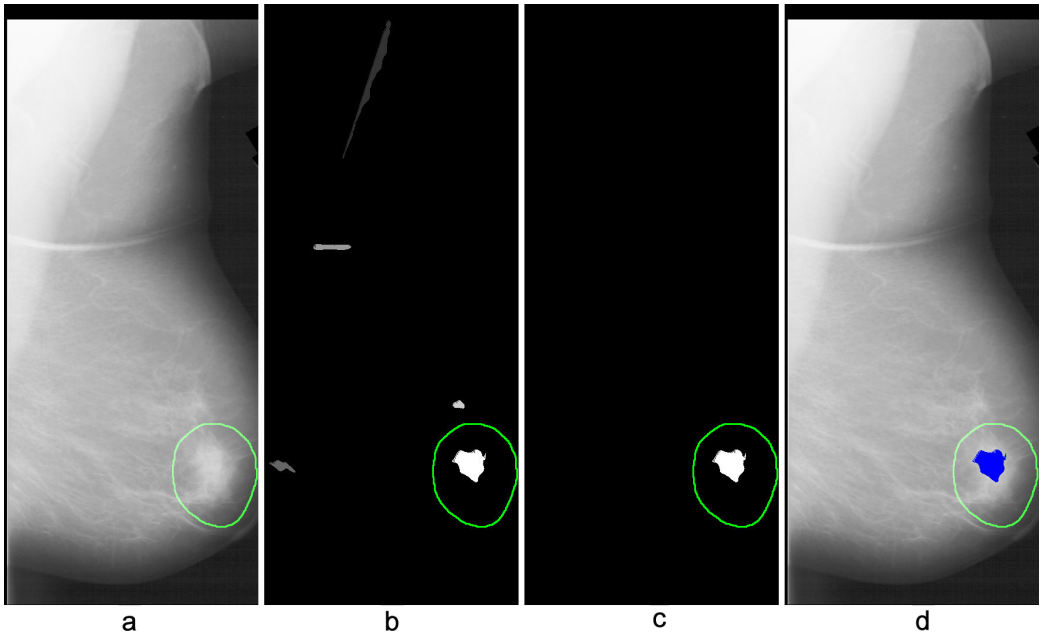


Figura 4.3: Estudo de caso para a imagem A1309 RIGHT MLO: (a) imagem original, (b) após a detecção de regiões suspeitas, (c) resultado da primeira redução de falso positivos, (d) resultado final após a segunda redução de falso positivos.

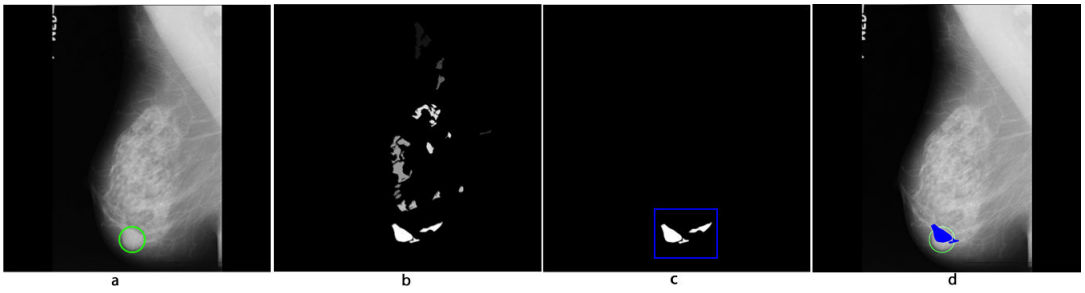


Figura 4.4: Estudo de caso para a imagem mdb021: (a) imagem original, (b) após a detecção de regiões suspeitas, (c) resultado da primeira redução de falso positivos, (d) resultado final após a segunda redução de falso positivos.

Problemas com a União

Embora os resultados analisados para a junção sejam interessantes, certos casos requerem atenção como o extremo apresentado pela Figura 4.6 e também pela Figura 4.7 onde a junção uniu áreas de não massa com a massa.

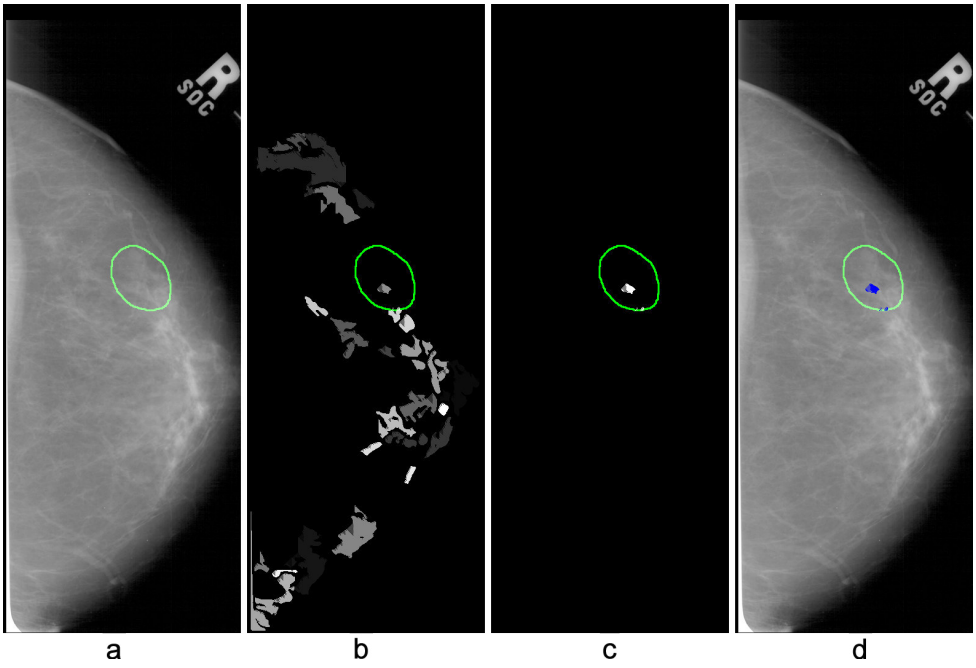


Figura 4.5: Estudo de caso para a imagem A1405 RIGHT CC: (a) imagem original, (b) após a detecção de regiões suspeitas, (c) resultado da primeira redução de falso positivos, (d) resultado final após a segunda redução de falso positivos.

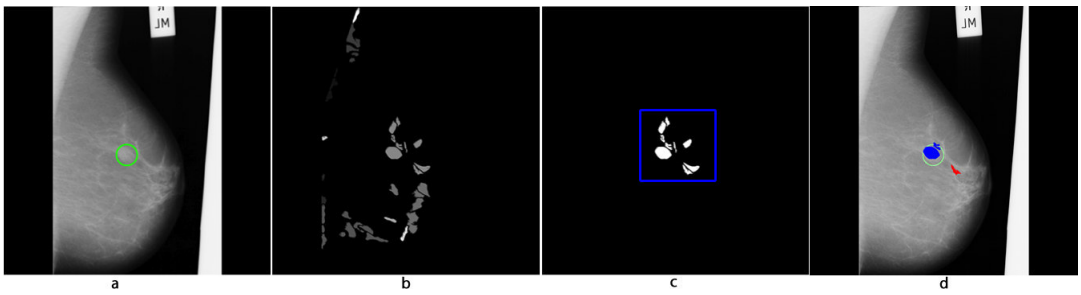


Figura 4.6: Estudo de caso para a imagem mdb012: (a) imagem original, (b) após a detecção de regiões suspeitas, (c) resultado da primeira redução de falso positivos e em (d) resultado final após a segunda redução de falso positivos.

São para estas situações que a metodologia propõe a utilização de uma segunda etapa de redução de falso positivos. Nesta segunda etapa, a quantidade de regiões já fora reduzida substancialmente. Assim, embora utilizando os mesmos métodos de descrição, o classificador terá a disposição uma base de informação mais restrita

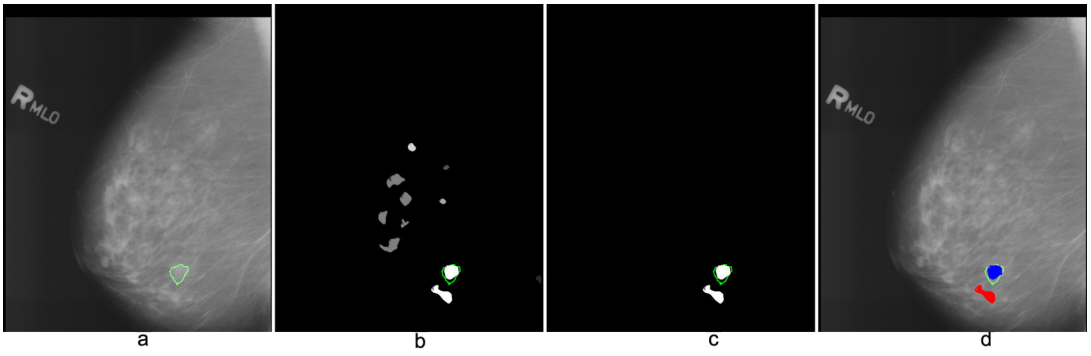


Figura 4.7: Estudo de caso para a imagem B3356 RIGHT MLO: (a) imagem original, (b) após a detecção de regiões suspeitas, (c) resultado da primeira redução de falso positivos e em (d) resultado final após a segunda redução de falso positivos.

e precisa, permitindo a classificação precisa de massas. Os resultados da segunda redução de falso positivos são evidenciados pela Figura 4.6(d) que apresenta o resultado final da metodologia. Nesta é possível verificar em azul as 2 regiões de massa detectadas (super segmentação) e em vermelho a única região falso positivo informada. Neste caso, a quantidade de falso positivos anotada foi três vezes maior do que a média para todas as imagens, evidenciando que na maior parte das imagens, nenhuma região de falso positivos foi gerada.

Para o exemplo apresentado pela Figura 4.7, a união agregou uma região normal a massa que não foi retirada após a segunda etapa de redução de falso positivos. A baixa taxa de união de regiões normais com massas foi observada nos testes com o DDSM. Esta é 1 das 7 imagens que tiveram falso positivos presentes após o final da metodologia para o DDSM.

Casos de falhas

Por fim, o último estudo de caso se remete às duas imagens em que não foram detectadas as massas, apresentadas pelas Figuras 4.8, 4.9 e 4.10. Na mdb126 verifica-se a presença de uma massa praticamente imperceptível, perdida durante a etapa filtro de densidade do FSA.

Na mdb080 a massa não possui alta densidade, mas após o filtro STD fica reduzida a uma pequena área por possui alta variação em sua borda e acaba por ser eliminada na etapa de filtro de área.

O caso apresentado para Figura 4.10 exemplifica os casos de erro que aconteceram na base DDSM, idênticos ao dois casos do MIAS: densidade baixa e região de massa demasiadamente pequena. Ambas as imagens são de densidade granular e possuem restrições não atendidas pela metodologia.

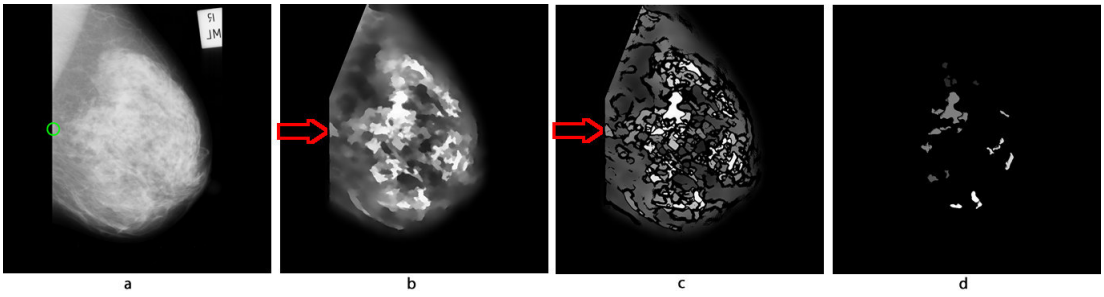


Figura 4.8: Estudo de caso para a imagem mdb126: (a) imagem original, (b) após melhoramento e em (c) após Filtro STD (d) após FSA.

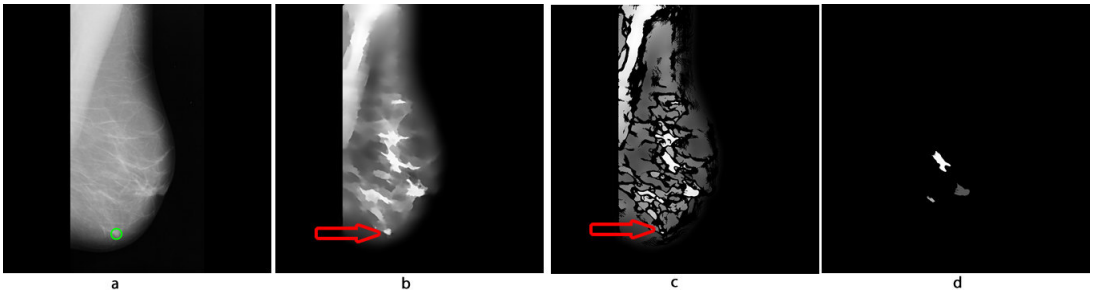


Figura 4.9: Estudo de caso para a imagem mdb080: (a) imagem original, (b) após melhoramento e em (c) após Filtro STD (d) após FSA.

4.4 Resumo de Resultados

Em geral verifica-se que para as duas bases testadas, os resultados de sensibilidade e taxa média de falso positivos são promissores. O melhor resultado para a base MIAS obteve sensibilidade equivalente a 97,30 com 0,3333 falso positivos médio por imagem e AFROC equivalente a 0,89 como demonstrado pela Figura 4.11.

O melhor resultado usando a base DDSM obteve sensibilidade de 91,63% com

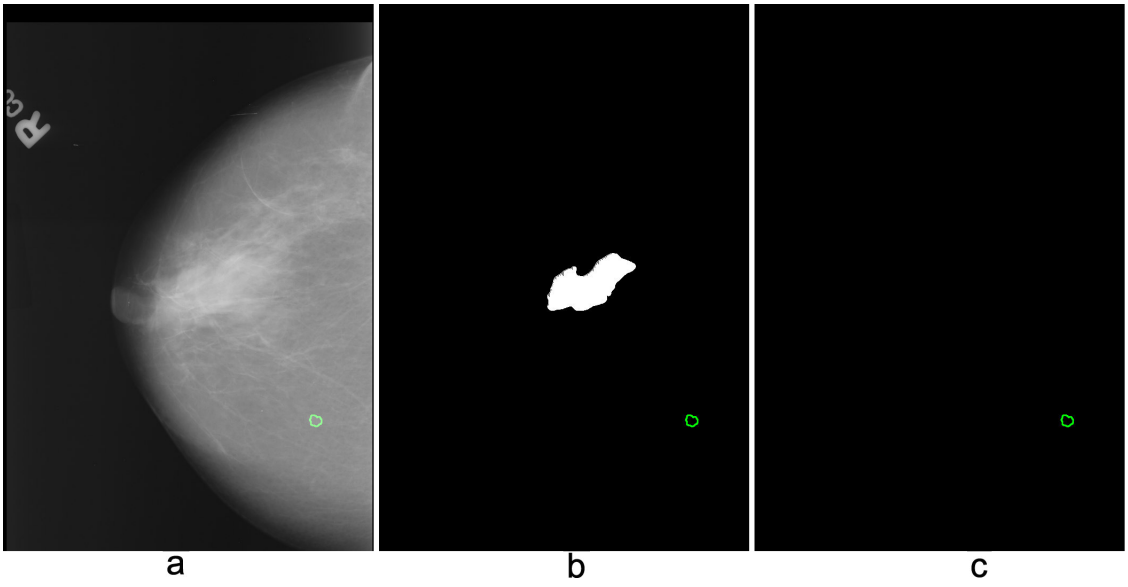


Figura 4.10: Estudo de caso para a imagem B3098 RIGHT CC: (a) imagem original, (b) após o FSA e em (c) após a primeira redução de falso positivos

0,013 falso positivos médio por imagem. A Figura 4.12 apresenta a curva FROC para o melhor resultado usando a base DDSM, com AFROC equivalente a 0,86.

O resultado é comparável aos melhores resultados apresentados nos trabalhos relacionados (Tabela 4.4) com a vantagem de não deteriorar a sensibilidade ao mesmo tempo que reduz o erro geral.

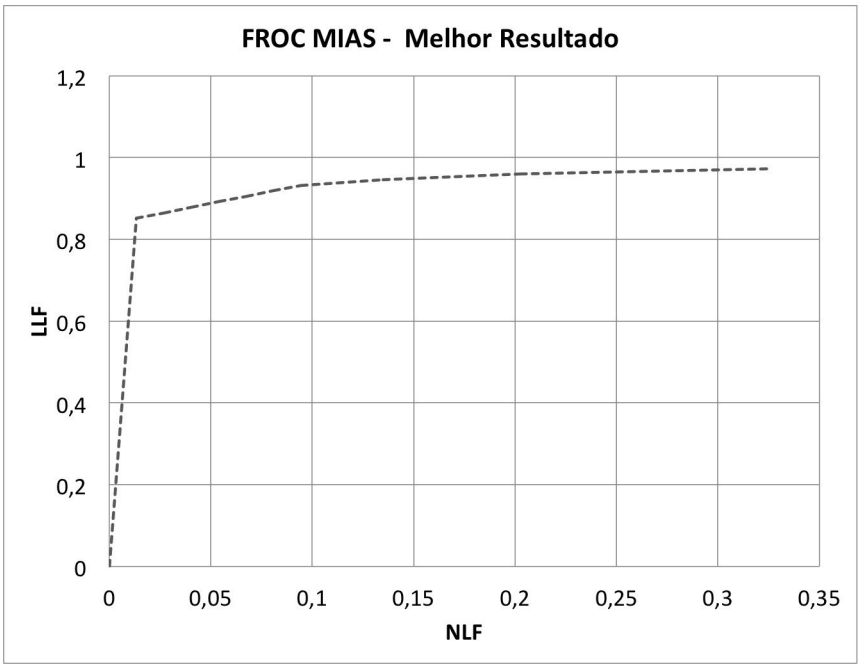


Figura 4.11: Curva FROC para a base MIAS, usando na primeira redução de falso positivos Geometria Côncava com 8 subpopulações, e na segunda redução de falso positivos Geometria Côncava com 3 subpopulações

Tabela 4.15: Resultados obtidos na segunda redução de falso positivos usando geometria côncava para a base DDSM.

Base	Subpopulação	$S_{img}(\%)$	FP/i	Acc (%)	nSv
3	7	91,63	0,000	100,00	1056
3	6	91,63	0,000	100,00	1067
6	9	91,63	0,007	99,63	641
6	10	91,63	0,008	99,54	727
4	12	91,63	0,013	99,26	365
4	11	91,63	0,013	99,26	797
6	7	91,63	0,016	99,17	481
4	10	91,63	0,016	99,17	723
6	12	91,63	0,016	99,17	877
3	12	91,63	0,016	99,17	1007
3	11	91,63	0,037	98,07	312
6	11	91,63	0,037	98,07	312
3	10	91,63	0,039	97,98	297
4	9	91,63	0,046	97,61	266
6	5	91,63	0,063	96,69	170
3	4	91,63	0,067	96,50	138
4	4	91,63	0,067	96,50	138
3	3	91,63	0,077	95,95	182
4	3	91,63	0,084	95,58	140
4	8	91,63	0,084	95,58	191
4	7	91,63	0,084	95,58	503
3	5	91,63	0,086	95,49	144
3	9	91,63	0,086	95,49	156
6	3	91,63	0,088	95,40	149
6	4	91,63	0,093	95,12	127
3	8	91,63	0,095	95,03	157
6	8	91,63	0,095	95,03	157
4	5	91,63	0,097	94,94	132
4	6	91,63	0,097	94,94	142
6	6	91,63	0,097	94,94	145

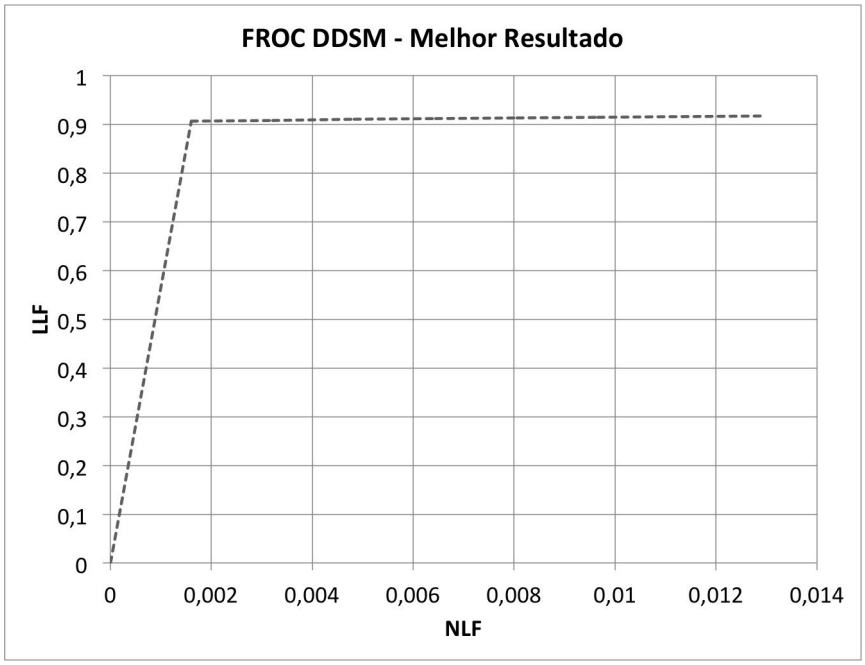


Figura 4.12: Curva FROC para a base DDSM, usando na primeira redução de falso positivos Geometria Côncava com 4 subpopulações, e na segunda redução de falso positivos Geometria Côncava com 12 subpopulações

Tabela 4.16: Comparação do desempenho das metodologias de detecção e redução de falso positivos apresentadas na seção de trabalhos relacionados.

Trabalho	Base	Acc	S	Sp	ROC	AFROC	FP/i
(ENGELAND; KARSSEMEIJER, 2007)	Privada	–	61,00	–	–	–	0,1
(QIAN *et al.*, 2007)	Privada	–	91,00	–	–	–	0,875
(ELTONSY *et al.*, 2007)	DDSM	–	81,00	–	–	–	0,6
(KOM *et al.*, 2007)	Privada	–	95,91	–	–	–	0,033
(WU *et al.*, 2007)	Privada	–	85,00	–	–	–	1,15
(LLADÓ *et al.*, 2009)	DDSM	93,00	–	–	0,94	–	–
(MASOTTI *et al.*, 2009)	DDSM	–	70,00	–	–	–	0,92
(MOAYEDI *et al.*, 2010)	MIAS	91,52	–	–	–	–	–
(OLIVER *et al.*, 2010)	Privada	–	86,70	–	–	0,946	2
(GAO *et al.*, 2010)	DDSM	–	95,3	–	–	–	2,83
(HONG; SOHN, 2010)	DDSM	–	90,00	–	–	–	2,3
(ZHENG, 2010)	Privada	–	93,00	–	–	–	1,19
(MAZUROWSKI *et al.*, 2011)	Privada	71,90	61,50	–	–	–	2
(KE *et al.*, 2010)	MIAS	–	85,11	–	–	–	1,44
(SAHBA; VENETSANOPOULOS, 2010a)	MIAS	–	88,00	–	0,86	–	2,1
(SAHBA; VENETSANOPOULOS, 2010b)	MIAS	–	90,00	–	0,88	–	1,9
(TERADA *et al.*, 2010)	Privada	–	81,2	–	–	–	5,0
(WEI *et al.*, 2011a)	Privada	–	87,00	–	–	–	1
(LIU *et al.*, 2011)	DDSM	–	76,8	–	–	–	1,36
(DEEPAK *et al.*, 2012)	MIAS	98,90	100,00	97,00	0,98	–	–
(WANG *et al.*, 2012)	Privada	–	84,00	–	–	–	0,69
(RAHMATI *et al.*, 2012)	DDSM	–	86,85	–	–	–	–
(TAI *et al.*, 2013)	DDSM	–	90,3	–	–	–	4,8
Metodologia Proposta	MIAS	–	97,30	–	–	0,89	0,33
Metodologia Proposta	DDSM	–	91,63	–	–	0,86	0,013

Capítulo 5

Conclusão

Este documento apresenta o desenvolvimento de uma metodologia para detecção de massas através de técnicas de processamento de imagens e reconhecimento de padrões. Nesse contexto, foi apresentado a fundamentação teórica dos principais assuntos relacionados a área de estudo, os quais são necessários para a compreensão do trabalho desenvolvido e também das técnicas abordadas na metodologia. Também apresentamos alguns trabalhos relacionados nas área de detecção de massas e também de redução de falso positivos.

A metodologia proposta foi validada através de testes usando as bases MIAS e DDSM. Analisando os resultados de detecção, é possível verificar que a metodologia pode ser adaptada para diferentes bases de imagens bastando o ajuste dos parâmetros de filtragem de regiões suspeitas.

Os melhores resultados foram obtidos, utilizando as duas bases de mamografias, quando se usou geometria côncava como um passo anterior para descrição de formas baseadas em texturas internas às regiões analisadas. As sensibilidades foram de 97,30% e 91,63% para as bases MIAS e DDSM, respectivamente. As taxas médias de falso positivos foram de 0,333 e 0,013 para MIAS e DDSM. Os resultados demonstram que formas côncavas de massas, baseadas em sua distribuição de textura interna, são um importante fator discriminador a ser usado para na redução de falso positivos.

Para os melhores resultados obtidos, verifica-se uma redução significativa de falso positivos em relação a etapa inicial de detecção (91,8% quando usando o MIAS e 99,65% quando usando o DDSM) mantendo a taxa de sensibilidade

constante, objetivo primário deste trabalho. Mesmo antes da redução de falso positivos, a taxa média de falso positivos gerada pode ser considerada mediana (se comparado com outros trabalho, conforme Tabela 4.4), fato que corrobora com a suposição de que a segmentação de massas em mamografias baseada em densidade é uma estratégia que atingiu neste trabalho seus objetivos de sensibilidade.

Analisando o desempenho da utilização de abordagens de zoneamento para o tratamento de índices de diversidade e o tratamento da característica local em índices geoestatísticos. Ambas abordagens consistem como maneiras de melhoria da eficiência da descrição sob o conjunto de diferentes visões geradas pelas várias decomposições espaciais e de distribuição de textura. Suas eficiências são comprovadas pelo resultado obtido quando aplicadas, tornando os índices de diversidade e geoestatísticos quase tão discriminatórios quanto o melhor resultado geral, se analisarmos a maioria dos testes realizados.

Em termos de tempo de execução, a metodologia pode ser analisada sobre duas posições, tempo de ajuste/treinamento e tempo de teste para um caso. O primeiro envolve toda a geração dos modelos de ajustamento de parâmetros para segmentação, extração de características e modelo de treinamento MVS. Este é caracterizado pelo maior tempo de execução, sendo em média necessários 4 minutos por imagem (tomando como base um computador i7, 2 núcleos, 2.3 Ghz, 8gb RAM DDR3). Leva-se também em consideração que este tempo está somado todas as análises de extração de características, fato que numa situação prática não será usado. O tempo de teste para uma imagem é em média de 30 segundos (tomando como base o mesmo computador relacionado acima). Logo, a metodologia gera um tempo de treinamento alto e que cresce em função da quantidade de imagens usadas como base. Mas o tempo de teste é relativamente baixo, validando a situação de aplicação prática da mesma.

Finalmente, esta tese apresenta como principais contribuições:

- Implementação de uma metodologia de detecção de massas baseada em densidade, diversidade e formas côncavas de textura;
- Proposição de uma abordagem espacial para utilização de índices de diversidade com intuito de realçar a extração de características;
- Utilização de índices conhecidos em outras áreas de pesquisa para extração

de textura em regiões extraídas de mamografias e redução de falso positivos:

  - Utilização dos índices de Diversidade Total, Berger-Parker, equitabilidade ED, Hill e Camargo;
  - Utilização dos índices geoestatísticos ripley e moran em sua abordagem local;
  - Utilização dos índices NNA e Joint Count;

- Uso de geometria côncava através de *Alpha-Shapes* como mecanismo de representação de intra formas descritas pela textura de uma região extraída da mamografia;
- Proposição das medidas de verificação direta de densidade em cascas extraídas pelo *Alpha-Shapes*:
  - Densidade Quadrangular, Densidade Anular, Densidade Quadrática

Portanto, a metodologia proposta por este trabalho cumpre os objetivos motivadores da mesma de apresentar uma nova abordagem que tem como principal característica a capacidade de ser implementada em situações reais devido aos resultados obtidos em termo de sensibilidade e taxa média de falso positivos. Estes que por sua vez, se posiciona bem quando comparados com os trabalhos recentes. Vale ainda analisar a robustez da mesma ao ser testada em duas bases de mamografias públicas.

5.1 Trabalhos Futuros

O tema abordado por esta tese tem sido pesquisado por vários grupos em função da construção de uma ferramenta a ser distribuída para o especialista. Algumas ferramentas semelhantes já foram desenvolvidas. Logo, um dos trabalhos futuros é a construção da ferramenta usando a metodologia proposta nesse trabalho, e assim, disponibilizar uma ferramenta CAD para distribuição.

Além da construção da ferramenta CAD, é necessário o aprimoramento da metodologia através de:

1. Construção de um MVS que suporte o treinamento incremental e evitar a sobrecarga realizada pela abordagem *Leave-One-Image-Out*;

2. Construção de uma arquitetura hierárquica de classificação, que utilize uma combinação de modelos de extração de características para realizar a classificação usando a mesma ferramenta de classificação;

3. Estudar a importância individual das características geradas por cada abordagem e identificar o melhor subconjunto geral;

4. Construir medidas de concavidade para análise da geometria côncava com intuito de classificar massas quanto ao caráter de malignidade;

5. Adotar uma análise individualizada de contornos da geometria côncava (evitar a média dos contornos) e construir um mecanismo que alterne os contornos utilizados para o reconhecimento de padrões, de acordo com aquele que maximize a acurácia;

6. Construir mecanismos de detecção hierárquico de massas baseado em contornos côncavos usando como base o padrão gerado pelo conjunto de contornos e informação de concavidade;

7. Analisar o desempenho da detecção após o registro das imagens da mamografia de maneira bilateral ou ipsilateral, com intuito de aumentar a capacidade de distinguir regiões de não massa;

8. Testar a metodologia para realizar a etapa seguinte a detecção que consiste no diagnóstico da região detectada quanto ao aspecto de malignidade; e

9. Estender a metodologia para a detecção de outras anormalidades da mama: calcificações e distorção arquitetural.

Referências

ACR. *Breast Imaging Reporting and Data System: Breast Imaging Atlas*. USA, 2003.

ACR, B.-R. C. of American College of R. *Bi-Rads: Atlas - Mammography*. Reston, VA: Preston White Drive, American College of Radiology, 2003.

ALTMAN, D.; BLAND, J. Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal*, v. 308, n. 6943, p. 1552–1562, 1994.

ANSELIN, L. Computing Enviroments for Spatial Data Analysis. *Journal of Geographical Systems*, v. 2, p. 201–220, 2001.

BASHEER, N. M.; MOHAMMED, M. H. Segmentation of breast masses in digital mammograms using adaptive median filtering and texture analysis. *Int. J. Recent Technol. Eng.(IJRTE)*, v. 2, n. 1, p. 39–43, 2013.

BIRD, R.; WALLACE, T.; YANKASKAS, B. Analysis of cancers missed at screening mammography. *Radiology*, Radiological Society of North America, v. 184, n. 3, p. 613–617, 1992.

BISHOP, C. *Pattern recognition and machine learning*. Secaucus, NJ: Springer-Verlag, 2006.

BORNEFALK, H.; HERMANSSON, A. On the comparison of froc curves in mammography cad systems. *Medical physics*, v. 32, p. 412–422, 2005.

BOYLE, P.; LEVIN, B. *et al. World cancer report 2008.* Lyon, France: IARC Press, International Agency for Research on Cancer, 2008.

BRAZ JUNIOR, G.; PAIVA, A. Cardoso de; SILVA, A. C.; OLIVEIRA, A. Cesar Muniz de. Classification of breast tissues using moran's index and geary's coefficient as texture signatures and svm. *Computers in Biology and Medicine*, Elsevier, v. 39, n. 12, p. 1063–1072, 2009.

BRAZ JUNIOR, G.; ROCHA, S. V. da; GATTASS, M.; SILVA, A. C.; PAIVA, A. C. d. A mass classification using spatial diversity approaches in mammography images for false positive reduction. *Expert Systems with Applications*, Pergamon, v. 40, n. 18, p. 7534–7543, 2013.

BUZAS, M.; HAYEK, L. She analysis for biofacies identification. *The Journal of Foraminiferal Research*, CFFR, v. 28, n. 3, p. 233–239, 1998.

CÂMARA, G. *Análise Espacial de Dados Geográficos*. 2003. Homepage do tutorial sobre "Análise Espacial", apresentado na Escola de Verão do IMPA (1999) e nos congressos GIS Brasil (1999, 2000 e 2001) e GeoBrasil (2000, 2001, 2002 e 2003).

CAMARGO, J. Must dominance increase with the number of subordinate species in competitive interactions. *Journal of Theoretical Biology*, Elsevier, v. 161, n. 4, p. 537–542, 1993.

CANNY, J. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, n. 6, p. 679–698, 1986.

CHAKRABORTY, D. P. *What is an FROC curve*. 2014.

CHENG, Y. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, v. 17, n. 8, p. 790–799, 1995.

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

DEEPAK, K. S.; MEDATHATI, N. K.; SIVASWAMY, J. Detection and discrimination of disease related abnormalities based on learning normal cases. *Pattern Recognition*, Elsevier, v. 45, p. 3707–3716, 2012.

DING, J.; KUO, C.; HONG, W. An efficient image segmentation technique by fast scanning and adaptive merging. *Graphical Models and Image Processing*, 2009.

DUARTE, D. L. *A Mama em Imagens*. Rio de Janeiro: Guanabara/Koogan, 2006.

ELTONSY, N.; TOURASSI, G.; ELMAGHRABY, A. A concentric morphology model for the detection of masses in mammography. *Medical Imaging, IEEE Transactions on*, IEEE, v. 26, n. 6, p. 880–889, 2007.

ENGELAND, S. van; KARSSEMEIJER, N. Combining two mammographic projections in a computer aided mass detection method. *Medical Physics*, v. 34, p. 898–905, 2007.

FREER; ULISSEY, M. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology*, v. 220, n. 3, p. 781–786, 2001. ISSN 0033-8419.

GAO, X.; WANG, Y.; LI, X.; TAO, D. On combining morphological component analysis and concentric morphology model for mammographic mass detection. *Information Technology in Biomedicine, IEEE Transactions on*, IEEE, v. 14, n. 2, p. 266–273, 2010.

GONZALEZ, R.; WOODS, R. *Processamento Digital de Imagens*. 3. ed. São Paulo: Pearson Prentice Hall, 2010.

GUR, A. I. B.; ROCKETTE, H. E.; SONG, T.; DAVID. Area under the free-response roc curve (froc) and a related summary index. *Journal of International Biometric Society*, v. 65, n. 1, p. 247–256, 2009.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. H. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, IEEE, n. 6, p. 610–621, 1973.

HAYKIN, S.; ENGEL, P. *Redes Neurais: Principios e Pratica*. Rio de Janeiro: Bookman, 2001.

HEATH, M.; BOWYER, K.; D., K. Current status of the digital database for screening mammography. *Digital Mammography*, v. 1, p. 457–460, 1998.

HILL, M. O. Diversity and evenness: a unifying notation and its consequences. *Ecology*, v. 54, p. 427–432, 1973.

HONG, B.-W.; SOHN, B.-S. Segmentation of regions of interest in mammograms in a topographic approach. *Information Technology in Biomedicine, IEEE Transactions on*, IEEE, v. 14, n. 1, p. 129–139, 2010.

INCA. Estimativas 2014: Incidência de Câncer no Brasil. Http://www.inca.gov.br/estimativa/2014/. 2014.

JOST, L. The relation between evenness and diversity. *Diversity*, Molecular Diversity Preservation International, v. 2, n. 2, p. 207–232, 2010.

KE, L.; MU, N.; KANG, Y. Mass computer-aided diagnosis method in mammogram based on texture features. In: IEEE. *3rd International Conference on Biomedical Engineering and Informatics (BMEI)*. Yantai, China, 2010. v. 1, p. 354–357.

KOM, G.; TIEDEU, A.; KOM, M. Automated detection of masses in mammograms by local adaptive thresholding. *Computers in Biology and Medicine*, Elsevier, v. 37, n. 1, p. 37–48, 2007.

KOPANS, D. B. *Breast Imaging*. Lippincott: Williams and Wilkins, 2007.

LANCASTER, J.; DOWNES, B. J. Spatial Point Pattern Analysis of Available and Exploited Resources. *Ecography*, Blackwell Synergy, v. 27, n. 1, p. 94–102, 2004.

LEVINE, N. *Análise Estatística de Dados Geográficos*. São Paulo, Brasil: Editora Unsep, 1996.

LIU, X.; XU, X.; LIU, J.; FENG, Z. A new automatic method for mass detection in mammography with false positives reduction by supported vector machine. In: *4th International Conference on Biomedical Engineering and Informatics*. Shangai, China: IEEE, 2011. v. 1, p. 33–37.

LLADÓ, X.; OLIVER, A.; FREIXENET, J.; MARTÍ, R.; MARTÍ, J. A textural approach for mass false positive reduction in mammography. *Computerized Medical Imaging and Graphics*, Elsevier, v. 33, n. 6, p. 415–422, 2009.

MACQUEEN, J. *et al.* Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. California, USA: University of California Berkeley, 1967. v. 1, n. 281-297, p. 14.

MAGURRAN, A. E. Measuring biological diversity. Taylor & Francis, 2004.

MASOTTI, M.; LANCONELLI, N.; CAMPANINI, R. Computer-aided mass detection in mammography: False positive reduction via gray-scale invariant ranklet texture features. *Medical physics*, v. 36, p. 311, 2009.

MAY, R. Patterns of species abundance and diversity. *Ecology and evolution of communities*, Harvard University Press, p. 81–120, 1975.

MAZUROWSKI, M.; LO, J.; HARRAWOOD, B.; TOURASSI, G. Mutual information-based template matching scheme for detection of breast masses: From mammography to digital breast tomosynthesis. *Journal of Biomedical Informatics*, Elsevier, v. 44, n. 5, p. 815–823, 2011.

MEERSMAN, D.; SCHEUNDERS, P.; DYCK, D. V. Detection of microcalcifications using non-linear filtering. In: *Signal Processing IX: theories and applications: proceedings of Eusipco*. Rhodes,Greece: Typorama Publications, 1998. v. 4, n. 1, p. 2465–2468.

MITCHELL, R. S.; PADWAL, R. S.; CHUCK, A. W.; KLARENBACH, S. W. *et al.* Cancer screening among the overweight and obese in canada. *American journal of preventive medicine*, v. 35, n. 2, p. 127, 2008.

MOAYEDI, F.; AZIMIFAR, Z.; BOOSTANI, R.; KATEBI, S. Contourlet-based mammography mass classification using the svm family. *Computers in Biology and Medicine*, Elsevier, v. 40, n. 4, p. 373–383, 2010.

MONTERO, R. S.; BRIBIESCA, E. State of the art of compactness and circularity measures. *International Mathematical Forum*, HIKARI P.O, v. 4, n. 25-28, p. 1305–1335, 2009.

MUCKE, H. E. Three-dimensional alpha shapes. *ACM Trans. Graph*, v. 13, p. 43–72, 1994.

OBUCHOWSKI, N. Roc analysis. *American Journal of Roentgenology*, v. 184, p. 364–372, 2005.

OLIVER, A.; LLADÓ, X.; FREIXENET, J.; MARTÍ, R.; PÉREZ, E.; PONT, J.; ZWIGGELAAR, R. Influence of using manual or automatic breast density information in a mass detection cad system. *Academic radiology*, v. 17, n. 7, p. 877–883, 2010.

PIELOU, E. *Ecological diversity*. New York: Wiley New York, 1975.

PIZER, S. M. Adaptive histogram equalization and its variotions. *Computer Vision, Graphics and Image Processing*, p. 355–368, 1987.

QIAN, W.; SONG, D.; LEI, M.; SANKAR, R.; EIKMAN, E. *et al.* Computer-aided mass detection based on ipsilateral multiview mammograms. *Academic radiology*, v. 14, n. 5, p. 530–538, 2007.

RAHMATI, P.; ADLER, A.; HAMARNEH, G. Mammography segmentation with maximum likelihood active contours. *Medical Image Analysis*, Elsevier, 2012.

RAMOS, R.; NASCIMENTO, M.; PEREIRA, D. Texture extraction: An evaluation of ridgelet, wavelet and co-occurrence based methods applied to mammograms. *Expert Systems with Applications*, Elsevier, 2012.

RIPLEY, B. D. Modelling Spatial Patterns. *J. Roy. Statist. Soc*, p. 172 – 212, 1977.

RSNA, R. S. of N. A. *New Mammography Technology Effective in Detecting Breast Cancer*. 2008.

SAHBA, F.; VENETSANOPOULOS, A. Breast mass detection using bilateral filter and mean shift based clustering. In: SPRINGER. *Proceedings of the 2010 International Conference on Signal Processing and Multimedia Applications (SIGMAP)*. Athens, Greece, 2010. p. 88–94.

SAHBA, F.; VENETSANOPOULOS, A. Mean shift based algorithm for mammographic breast mass detection. In: IEEE. *17th IEEE International Conference on Image Processing (ICIP)*. Hong Kong, 2010. p. 3629–3632.

SAMPAIO, W.; DINIZ, E. M.; SILVA, A. C.; PAIVA, A. C.; GATASS, M. Detection of masses in mammogram images using cnn, geostatistic functions and svm. *Computers in Biology and Medicine*, v. 41, p. 653–664, 2011.

SHANNON, C. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, ACM, v. 5, n. 1, p. 3–55, 2001.

SIMPSON, E. Measurement of diversity. *Nature; Nature*, 1949.

SOCIETY, A. A. C. *Learn about breast cancer*. 2013.

SOUSA, J. R. F. da S.; SILVA, A. C.; PAIVA, A. C. de; NUNES, R. A. Methodology for automatic detection of lung nodules in computerized tomography images. *computer methods and programs in biomedicine*, Elsevier, v. 98, n. 1, p. 1–14, 2010.

SUCKLING, J.; PARKER, J.; DANCE, D.; ASTLEY, S.; HUTT, I.; BOGGIS, C.; RICKETTS, I.; STAMATAKIS, E.; CERNEAZ, N.; KOK, S. *et al.* The Mammographic Images Analysis Society Digital Mammogram Database. *Exerpta Medica International Congress Series*, v. 1069, p. 375–378, 1994.

TAI, S.; CHEN, Z.; TSAI, W. An automatic mass detection system in mammograms based on complex texture features. *IEEE Journal of Biomedical and Health Informatics*, Early Access Online, 2013. ISSN 2168-2194.

TERADA, T.; FUKUMIZU, Y.; YAMAUCHI, H.; CHOU, H.; KURUMI, Y. Detecting mass and its region in mammograms using mean shift segmentation and iris filter. In: IEEE. *International Symposium on Communications and Information Technologies (ISCIT)*. Tokyo, 2010. p. 1176–1179.

TZIKOPOULOS, S.; MAVROFORAKIS, M.; GEORGIOU, H.; DIMITROPOULOS, N.; THEODORIDIS, S. A fully automated scheme for mammographic segmentation and classification based on breast density and asymmetry. *Computer Methods and Programs in Biomedicine*, Elsevier, v. 102, n. 1, p. 47–63, 2011.

USYSTEMS. *The somo v Automated Breast Ultrasound INSIGHT System,*. 2013.

VAFAIE, H.; IMAM, I. F. Feature selection methods: genetic algorithms vs. greedy-like search. In: *Proceedings of International Conference on Fuzzy and Intelligent Control Systems*. Louisville: IEEE, 1994.

VALE, P. O processo de detecção de bordas de canny: Fundamentos, algoritmos e avaliação experimental. In: *Anais do Simpósio Brasileiro de Geomática*. Presidente Prudente: UNESP, 2002. p. 292–303.

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

VOMWEG, T. *Image Processing in Radiology: Current Applications: Computer-Aided Diagnosis: Clinical Applications in the Breast*. Berlim: Springer, 2008.

WANG, X.; LI, L.; XU, W.; LIU, W.; LEDERMAN, D.; ZHENG, B. Improving performance of computer-aided detection of masses by incorporating bilateral mammographic density asymmetry: an assessment. *Academic Radiology*, Elsevier, v. 19, n. 3, p. 303–310, 2012.

WEI, J.; CHAN, H.; ZHOU, C.; WU, Y.; SAHINER, B.; HADJIISKI, L.; ROUBIDOUX, M.; HELVIE, M. Computer-aided detection of breast masses: Four-view strategy for screening mammography. *Medical physics*, American Association of Physicists in Medicine, v. 38, n. 4, p. 1867–1876, 2011.

WEI, J.; CHAN, H.-P.; ZHOU, C.; WU, Y.-T.; SAHINER, B.; HADJIISKI, L. M.; ROUBIDOUX, M. A.; HELVIE, M. A. Computer-aided detection of breast masses: Four-view strategy for screening mammography. *Medical physics*, v. 38, p. 1867, 2011.

WHO. *World Health Statistics 2012*. France: World Health Organization, 2012.

WU, Y.; WEI, J.; HADJIISKI, L.; SAHINER, B.; ZHOU, C.; GE, J.; SHI, J.; ZHANG, Y.; CHAN, H. Bilateral analysis based false positive reduction for computer-aided mass detection. *Medical physics*, NIH Public Access, v. 34, n. 8, p. 3334, 2007.

YUAN, Y.; GIGER, M.; LI, H.; SUZUKI, K.; SENNETT, C. A dual-stage method for lesion segmentation on digital mammograms. *Medical physics*, v. 34, p. 4180, 2007.

ZHENG, Y. Breast cancer detection with gabor features from digital mammograms. *Algorithms*, v. 3, p. 44–62, 2010.