RENORBIO

Programa de Pós-Graduação em Biotecnologia

Método de Detecção de Câncer em Mamas Densas Utilizando Diagnóstico Auxiliado por Computador

Lúcio Flávio de Albuquerque Campos

São Luís–MA

2013

Método de Detecção de Câncer em Mamas Densas Utilizando Diagnóstico Auxiliado por Computador

Tese apresentada ao Programa de Pós-Graduação em Biotecnologia, como parte dos requisitos para obtenção do título de Doutor em Biotecnologia.

Orientador: Prof. Dr. Allan Kardec Duailibe Barros Filho

São Luís – 2013

BANCA DE AVALIAÇÃO

Prof. Dr. Allan Kardec Duailibe Barros Filho

Universidade Federal do Maranhão – UFMA
Orientador

Prof. Dra. Maria do Desterro Soares Brandão Nascimento

Universidade Federal do Maranhão – UFMA
Membro

Prof. Dr. Francisco das Chagas de Souza

Universidade Federal do Maranhão – UFMA
Membro

Prof. Dra. Alcione Miranda dos Santos

Universidade Federal do Maranhão – UFMA
Membro

Prof. Dr. Ewaldo Éder Carvalho Santana

Universidade Estadual do Maranhão – UEMA
Membro

A eles, meu porto seguro.

Filha, *Jayne da Rocha Campos*

Mãe, *Elaine de Albuquerque Campos*

Pai, *José de Ribamar Campos Sobrinho*

Vó, *Conceição de Maria Sousa Campos*

Irmã, *Lívia Flávia de Albuquerque Campos*

AGRADECIMENTOS

A Deus, minha rocha e minha fortaleza.

À minha família, que tem acreditado em mim e nos meus sonhos.

Ao Programa de Pós-Graduação em Biotecnologia – RENORBIO a todos os seus professores.

Ao orientador e amigo, Professor Allan Kardec Barros, que soube conduzir com infinita maestria, cada etapa de minha vida acadêmica.

A minha namorada, Aline Furtado, pelo apoio incondicional e pelo companheirismo.

Aos meus amigos do Laboratório de Processamento da Informação Biológica, PIB-UFMA: Ewaldo, Denner, André, Daniel, Luís Cláudio, Marcos, Áurea, Cristiane, Éder, Anderson.

Aos meus companheiros de viagens, em disciplinas fora do estado: Richard e Mônica.

Aos meus amigos da UEMA, que muito me apoiaram, desde o início desde doutorado: Mariano, Cícero, Mauro, Reinaldo, Diógenes, Fernando.

Aos meus coordenadores da UEMANet, Roberto Serra, Fátima Rios, João Augusto, Ilka Serra, e Eliza Flora. Muito obrigado pelo apoio incondicional.

Aos demais amigos da UEMANet: Ariana, Kilton, Willian, Bruno, Marcello, Hans, Hugo, Cosme, Carla, Zélia,Dario, Marina e tantos outros que não cabem aqui nesta página.

"Uma publicação científica pode ser comparada a um pequeno tijolo que depositamos na imensa parede da Ciência. É a nossa pequena contribuição ao conhecimento da humanidade."

Prof. Dr. Paulo C. Razuk

RESUMO

O câncer de mama continua sendo o tipo de câncer de maior incidência e mortalidade entre as mulheres. O melhor método de prevenção é o diagnóstico precoce, que é realizado com o auxilio da mamografia. Contudo, a mamografia não é eficaz quando a mama apresenta uma composição superior a 50 % de tecido fibroglandular, ou seja, de tecido denso. Estudos comprovam que a densidade mamária elevada é apontada como um fator de risco para o desenvolvimento da doença, e devido a isso novas técnicas de diagnóstico de câncer em pacientes com mamas densas estão sendo estudados. Esta tese propõe um método de diagnóstico precoce de câncer, em mamas densas, consideradas pela literatura de difícil rastreio e detecção, com o objetivo de aumentar as chances de cura da paciente, e diminuir os casos de mortalidade da doença. A metodologia empregada no trabalho utilizou a base de dados MIAS para teste, técnicas de equalização adaptativa e alargamento de contraste, na fase de segmentação, e análise de componentes independentes, máxima relevância - mínima redundância e máquinas de vetor de suporte, na etapa de classificação. Os testes foram realizados com 76 mamogramas de mamas em que o parênquima denso dificulta a detecção. A partir dos testes realizados, obteve-se média de acerto de 97.36 % na etapa de segmentação. Já na etapa de classificação foi encontrada uma média de acerto de 97,2% com sensibilidade de 81.88% e especificidade de 100%. Baseado nos resultados encontrados, considerando que o método foi realizado apenas em mamogramas de difícil detecção, pode-se considerar que o método obteve excelente desempenho, justificando o teste em bases de dados maiores, e futuramente viabilizando seu uso em hospitais e clinicas de radiologia.

Palavras-chave: Mamograma, Câncer de Mama, Densidade Mamária, Diagnóstico Auxiliado por Computador.

ABSTRACT

Breast Cancer remains the type of cancer with the largest incidence and mortality in women. The best method of prevention is early diagnosis, which is carried out with mammography. However, a mammogram is not effective when the breast has a composition of greater than 50% fibroglandular tissue, or dense tissue. Studies show that high breast density is identified as a risk factor for developing the disease, and because of this new diagnostic technique for cancer in patients with dense breasts are being studied. This thesis proposes a method for early diagnosis of cancer in dense breasts, considered in the literature as hard scanning and detection. The methodology applied in this work used MIAS database for tests, equalization adaptive of histogram and contrast stretching techniques for segmentation step, and independent component analysis maxima-relevance-minimal-redundance and support vector machine for classification step. The tests were carried out with 76 breast mammograms whose dense parenchyma's make detection difficult. From the tests, we obtained accuracy of 97.36% in the segmentation stage. Already in the classification stage was an accuracy of 97.2% with a sensitivity of 81.88% and specificity of 100%. Based on the results, considering that the method was performed only on mammograms difficult to detect, it can be considered that the method achieved excellent performance, justifying the test in larger databases, and eventually enabling their use in hospitals and radiology clinics.

Keywords: Mammogram, Breast Cancer, Breast Density, Computer Aided Diagnosis

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AUC	***Area Under Curve*** **(Área sob a Curva)**
BI-RADS	***Breast Imaging-Reporting and Data System*** **(Sistema de Dados e Relatórios em Imagens de Mama)**
BSS	***Blind Source Separation*** **(Separação Cega de Fontes)**
CAD	**Diagnóstico Auxiliado por Computador**
CC	**Crânio Caudal**
CDF	**Função de Distribuição Cumulativa**
CDI	**Carcinoma Ductal Invasivo**
CDIS	**Carcinoma Ductal *in Situ***
CLAHE	***Contrast Limited Adaptive Histogram Equalization*** **(Equalização Adaptativa de Histograma com Limitação de Contraste)**
CLI	**Carcinoma Lobular Invasivo**
FPI	**Falso Positivo por Imagem**
ICA	***Independent Component Analysis*** **(Análise de Componentes Independentes)**
INCA	**Instituto Nacional do Câncer**
LCIS	**Lobular Carcinoma *in Situ***
MLO	**Médio Lateral Oblíqua**
mRMR	**Máxima Relevância e Mínima Redundância**
ROC	***Receiver Operating Characteristic*** **(Característica de Operação do Receptor)**
ROS	**Regiões Suspeitas**
SADIM	**Sistema de Auxilio de Diagnóstico em Imagens Mamográficas**
SVM	***Support Vector Machine*** **(Máquina de Vetor de Suporte)**
MIAS	***Mammographic Institute Analisys Society*** **– Instituto de Análises Mamográficas**

LISTA DE VARIÁVEIS

l	**Número de níveis de cinza da imagem**
h	**Número total de *pixels* da imagem**
$p_{(r)}$	**Probabilidade do j-ésimo nível de cinza**
h_j	**Número de *pixels* cujo nível de cinza corresponde a j**
g_j	**Função de distribuição cumulativa (CDF)**
g	**Valor do *pixel* atualizado**
g_{min}	**Valor do menor *pixel***
α	***Clip limit* (Clip Limite)**
x_n	**Sinal aleatório**
s_n	**Componente independente aleatório**
a_n	**Coeficiente, considerados como características da imagem.**
$\mathbf{S}$	**Matriz de componentes independentes**
$\mathbf{A}$	**Matriz de coeficientes**
$\mathbf{X}$	**Matriz de Mistura**
v	**Vetor de dados**
c	**Vetor de classe (rótulo)**
$\phi(D,R)$	**Máxima relevância e Mínima Redundância**
H_n	**Hiperplano Linearmente Separável**
$K(w,\epsilon,\rho)$	**Operador Lagrangiano**

SUMÁRIO

1. Introdução

O câncer da mama é o tipo de câncer que mais acomete as mulheres em todo o mundo, tanto em países em desenvolvimento quanto em países desenvolvidos. Segundo dados mais atuais do INCA, no ano de 2008, cerca de 1,4 milhões de mulheres foram acometidas pela doença em todo o mundo, o que representa um total de aproximadamente 23% de todos os tipos de câncer. A Tabela 1 apresenta os valores intermediários no padrão de incidência e controle do câncer de mama, no ano de 2008 (INCA, 2013).

Tabela 1 – Taxas de incidência e mortalidade por câncer de mama, por 100 mil mulheres, em países selecionados, 2008

Região / País	Incidência		Mortalidade	
	Taxa Bruta	Taxa Padronizada	Taxa Bruta	Taxa Padronizada
Finlândia	151,1	86,6	31,3	14,7
Reino Unido	146,2	87,9	38,3	18,6
Espanha	97,6	61,0	26,6	12,8
Estados Unidos	115,5	76,0	25,6	14,7
Canadá	136,9	83,2	30,2	15,6
Austrália	126,5	84,8	25,6	14,7
Japão	70,3	42,7	18,1	9,2
Paraguai	39,6	51,4	13,2	17,1
Bolívia	18,4	24,0	5,8	7,6
Zâmbia	11,2	20,5	6,3	12,2
Brasil *	**43,7**	**42,3**	**12,9**	**12,3**
Brasil (dados oficiais) **	49,3	-	11,6	11,1

Fonte: INCA (2013)

Segundo o Instituto Nacional do Câncer (INCA), no Brasil, dos 518.510 casos da doença em 2012, 52.680 são de câncer de mama, representando um total de 27,9% de

todos os casos de câncer entre as mulheres, com um risco de 52 casos a cada 100 mil mulheres. E no Nordeste, são 8.970 casos, sendo 3.440 nas capitais.

As taxas de mortalidade por câncer da mama continuam elevadas no Brasil, muito provavelmente porque a doença ainda é diagnosticada em estágios avançados. Entretanto é considerado um câncer relativamente de bom prognóstico quando diagnosticado e tratado precocemente. A sobrevida média após cinco anos da descoberta do diagnóstico na população de países desenvolvidos tem apresentado um discreto aumento, cerca de 85%. E nos países em desenvolvimento, a sobrevida fica em torno de 60% (INCA,2013).

A mamografia é uma radiografia das mamas, que possibilita a detecção/diagnóstico precoce do câncer, por ser capaz de encontrar lesões na fase inicial, na ordem de poucos milímetros. É realizada através de um mamógrafo, em que a mama é comprimida para melhorar a resolução da imagem, e gerar melhores resultados, ou seja, diagnóstico mais preciso. As duas mamas são analisadas separadamente e a partir delas são geradas imagens de duas visões: Crânio-Caudal (CC) e Médio Lateral Oblíqua (MLO).

O sucesso da eficácia do exame depende de vários fatores que podem refletir na sua qualidade (densidade do tecido mamário, habilidade do radiologista, mamógrafos em perfeito estado). Devido a essa combinação de fatores, estima-se que o erro de diagnóstico de mamografia esteja na ordem de 30% (PAQUERAULT, 2009).

Dentre um destes fatores que podem refletir na eficácia do exame, destaca-se a densidade do tecido mamário. A mama gordurosa absorve uma menor quantidade de raios-X, aparecendo mais escura no exame de mamografia, enquanto uma mama densa – que possui uma proporção de tecidos fibroglandulares superior a 25% do tecido total- apresenta densidade óptica maior e aparecem tons mais claros. Tecidos lesionados e

calcificações aparecem em tonalidades mais claras na imagem obtida após a revelação do filme mamográfico, mas essa diferenciação fica prejudicada em imagens de mamas densas, o que dificulta o sucesso do exame (BOYD, 1995).

O erro de aproximadamente 30% do diagnóstico da mamografia motivou nos últimos 20 anos, o surgimento de pesquisas de ferramentas computacionais para auxiliar o médico/radiologista na decisão e suporte de diagnóstico. Tais ferramentas, chamadas de Diagnóstico Auxiliado por Computador (CAD) fornecem uma dupla leitura, ou seja, uma segunda opinião ao especialista, aumentando as taxas de acerto precoce, fazendo da mamografia uma ferramenta mais confiável (TANG et al, 2009).

Além da dificuldade de interpretação do exame, estudos recentes apontam que mulheres com densidade mamográfica superior a 75%, possuem o risco aumentado de desenvolver câncer de mama entre 4 a 6 vezes, comparado com mulheres que possuem pouco ou nenhum tecido denso na mama (BOYD, 2007).

Este trabalho propõe uma metodologia CAD para diagnóstico de câncer, específico para mamas densas, consideradas as de maior dificuldade de interpretação e de grande fator de risco. A metodologia utiliza equalização adaptativa de histograma com limitação de contraste para identificar as regiões suspeitas de lesão, chamadas de ROS – *Regions on Suspicious*. Em seguida, é utilizada a análise de componentes independentes, somada com o algoritmo de máxima relevância e mínima redundância (mRMR), para extrair características e encontrar o melhor conjunto das características obtidas, que finalmente serão analisadas por um classificador, baseado em máquina de vetor de suporte – SVM, para decidir se as regiões suspeitas de lesão são anormais ou normais. Nesse contexto, será considerada anormal qualquer região que corresponda a uma neoplasia, seja ela benigna ou maligna.

1.1 Organização do Trabalho

O Capítulo 2 aborda revisão de literatura necessária ao desenvolvimento da metodologia proposta. Apresenta alguns conceitos de processamento de sinais e imagens utilizados, tais como: equalização adaptativa com limitação de contraste; análise de componentes independentes; máxima-relevância e mínima-redundância; máquinas de vetor de suporte.

O Capítulo 3 mostra os objetivos gerais e específicos do trabalho.

O Capítulo 4 descreve o material utilizado e a metodologia proposta, divididas em: Pré-processamento, segmentação das regiões suspeitas, extração e seleção de características e a classificação das regiões suspeitas em anormais e normais.

O Capítulo 5 apresenta os resultados e discussões obtidos baseados na metodologia proposta.

O Capítulo 6 apresenta a conclusão sobre o trabalho, mostrando a eficiência da metodologia proposta e sugestões para trabalhos futuros.

2. REVISÃO TEÓRICA

2.1.O Câncer

Câncer é o nome geral dado a um conjunto de mais de 100 doenças, que têm em comum o crescimento descontrolado de células, que tendem a invadir tecidos e órgãos vizinhos.

As células normais que formam os tecidos do corpo humano são capazes de se multiplicar por meio de um processo contínuo que é natural. A maioria das células normais cresce, multiplica-se e morre de maneira ordenada.

O crescimento das células cancerosas é diferente do crescimento das células normais. As células cancerosas, em vez de morrerem, continuam crescendo incontrolavelmente, formando outras novas células anormais, que se dividem de forma rápida, agressiva e incontrolável, espalhando-se para outras regiões do corpo – acarretando transtornos funcionais.

O crescimento celular descontrolado causado pelo câncer resulta em uma massa anormal de tecido, que possui crescimento autônomo, persistindo dessa maneira após o término dos estímulos que o provocaram. As chamadas neoplasias (câncer *in situ* e câncer invasivo) correspondem a essa forma não controlada de crescimento celular e, na prática, são denominadas tumores. A Figura 1 ilustra células com crescimento descontrolado, originando cânceres *in situ* e invasivos (INCA, 2013).

As neoplasias são consideradas proliferações anormais do tecido, que fogem parcial ou totalmente ao controle do organismo, com efeitos agressivos sobre o homem.

As Neoplasias podem ser benignas ou malignas. As neoplasias benignas ou tumores benignos têm seu crescimento de forma organizada, geralmente lento, expansivo e apresentam limites bem nítidos. Apesar de não invadirem os tecidos vizinhos, podem comprimir os órgãos e tecidos adjacentes.

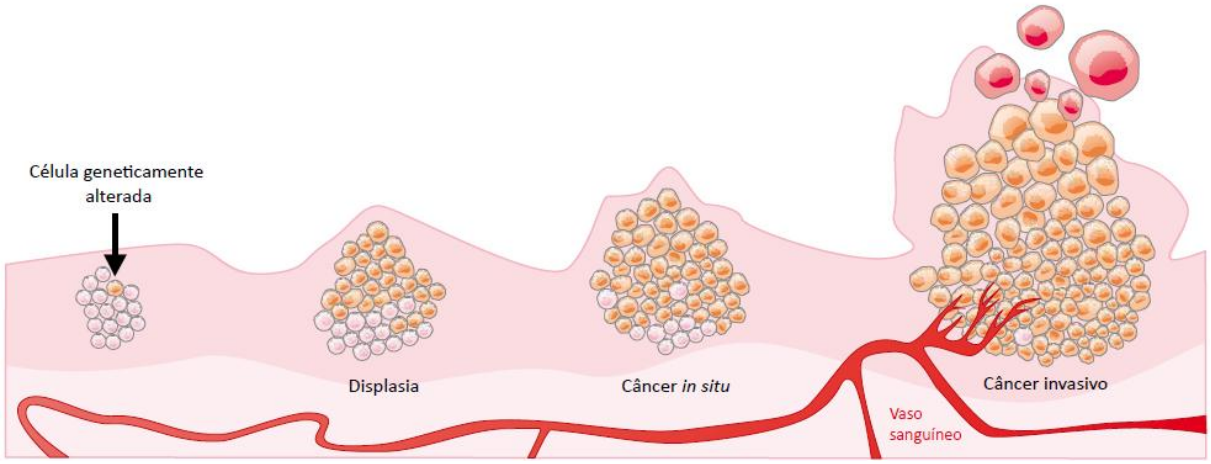


Figura 1 – Crescimento celular descontrolado, originando desde cânceres *in situ*, até cânceres invasivos, e posteriormente metástases. Reprodução: INCA, 2013

As neoplasias malignas ou tumores malignos manifestam um maior grau de autonomia e são capazes de invadir tecidos vizinhos e provocar metástases, podendo ser resistentes ao tratamento e causar a morte. A Figura 2 ilustra a diferença entre tumor benigno e maligno (INCA, 2013).

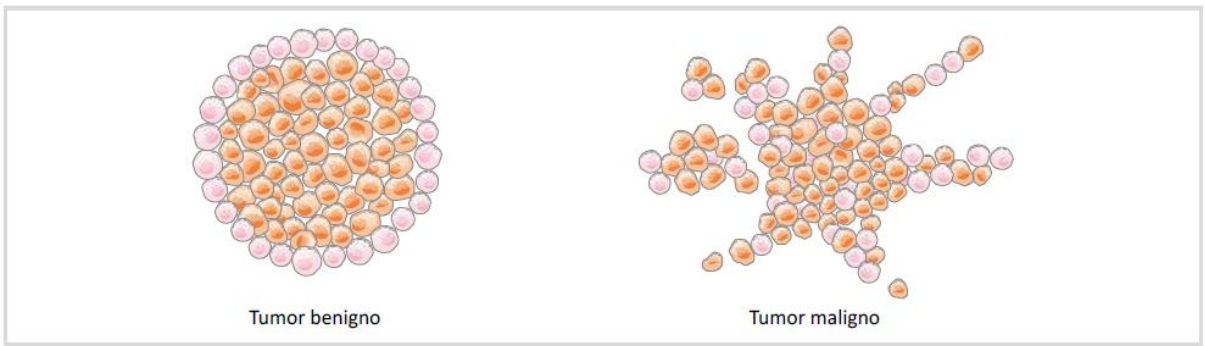


Figura 2 – Diferenças entre o tumor benigno e tumor maligno. Reprodução: INCA,2013.

2.2. As Glândulas Mamárias

Segundo Wexner (2013), as glândulas mamárias, como ilustra a Figura 3 estão situadas na parede anterior do tórax, tendo como principal função a secreção do leite e se compõem de:

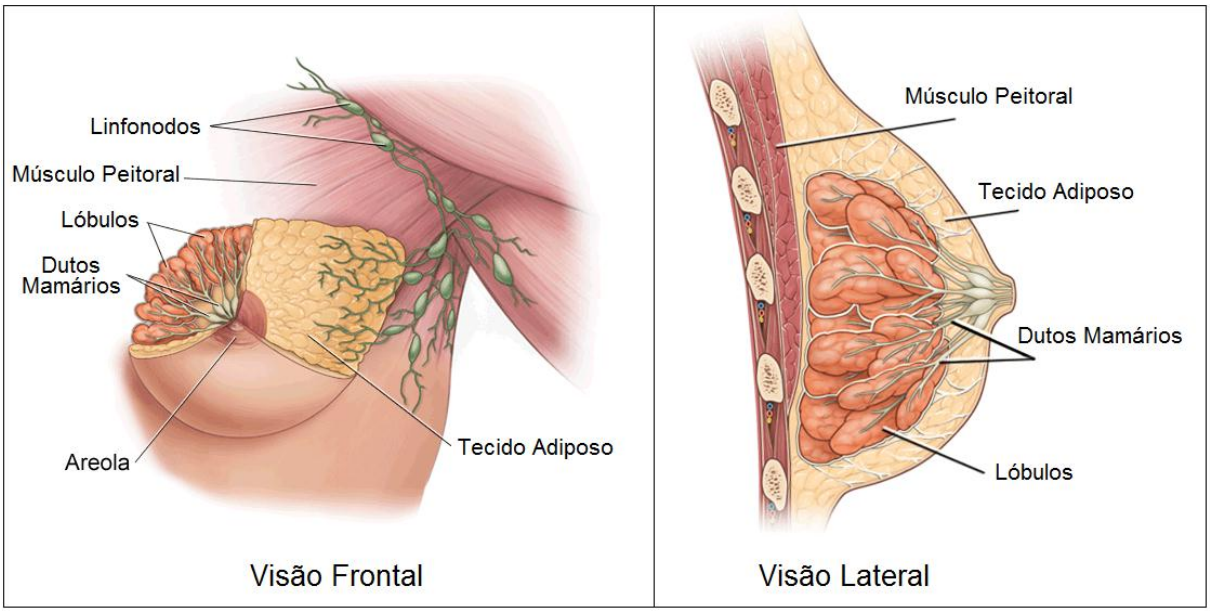


Figura 3 – Visão Frontal e Lateral da Mama. Reprodução (WEXNER,2013)

- Dutos Mamários: em número de 15 a 20 canais, conduzem a secreção (leite) até a papila;
- Lóbulos Mamários: conjunto de ácinos (menor parte da glândula, responsável pela produção de leite na lactação);
- Tecido Glandular: conjunto de lóbulos e dutos;
- Tecido Adiposo: todo o restante da mama é preenchido por tecido adiposo, cuja quantidade varia com as características físicas, estado nutricional e idade da mulher.

2.2.1. Tipos de Câncer de Mama

O câncer de mama pode se manifestar em vários tecidos da mama, tais como ductos, lóbulos, ou em tecidos entre ductos e lóbulos. Podemos classificar o câncer de mama em duas formas: cânceres não invasivos (*in situ)* e os cânceres invasivos ou infiltrantes. O primeiro, não invasivo ou *in situ*, apresenta células doentes que se originam dentro dos ductos ou dos lóbulos, que são estruturas que fazem parte da anatomia normal das mamas, mas não invadem ou infiltram estruturas próximas e nem são capazes de originar uma metástase. Já o último, o invasivo ou infiltrante, pode invadir tecidos próximos ou até mesmo órgãos distantes originando a metástase (BREASTCANCER, 2013).

Nas linhas abaixo, será descrito os tipos mais comuns de câncer *in situ* e invasivo.

2.2.1.1. Carcinoma Ductal *in situ*

O carcinoma ductal *in situ* (CDIS) é o tipo mais comum de câncer de mama não invasivo. É um câncer que ocorre nos ductos mamários, canal por onde passa o leite até os mamilos. CDIS é chamado de "não invasiva", porque não se espalhou além do ducto de leite em qualquer tecido mamário normal circundante.

O carcinoma ductal *in situ* é considerado uma lesão precursora, ou seja, se é deixado na mama, poderá, com grande probabilidade, evoluir para carcinoma invasivo.

O tratamento do carcinoma ductal *in situ* é a retirada total da lesão ou através de mastectomia (quando é muito extenso) ou cirurgia conservadora (quando é menor) seguido de radioterapia (BREASTCANCER,2013).

A Figura 4 ilustra um ducto mamário com CDIS

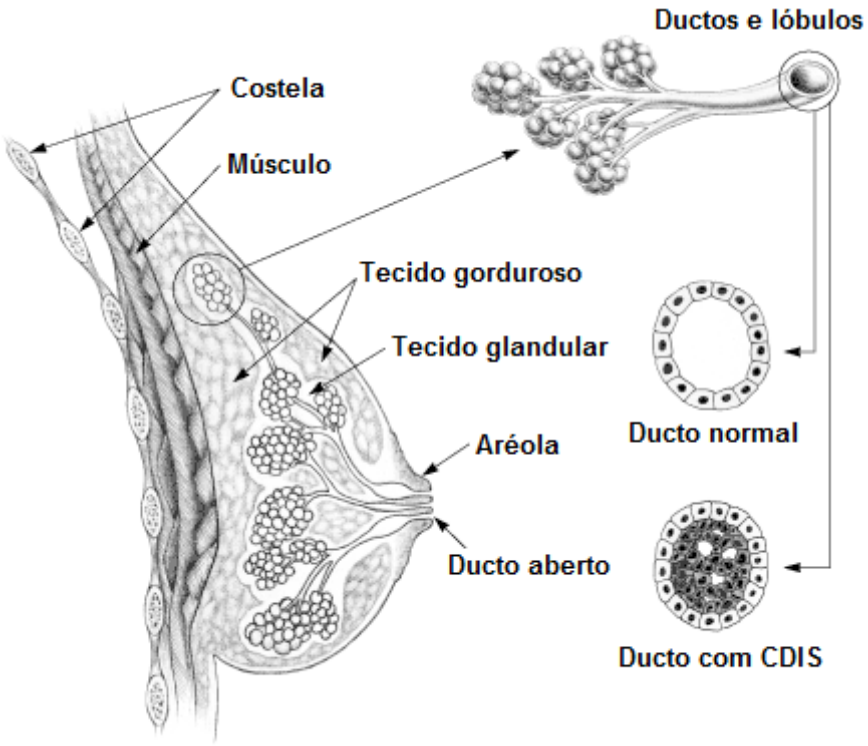


Figura 4 – Ilustração de uma mama com CDIS. Reprodução (CANCER COUNCIL, 2007)

2.2.1.2. Carcinoma Lobular *in situ*

O carcinoma lobular *in situ* (CLIS) ocorre nos lóbulos, glândulas produtoras de leite no final dos ductos mamários.

Apesar do fato de seu nome incluir o termo "carcinoma", CLIS não é um câncer mamário, e sim uma indicação de que a paciente possui uma probabilidade mais alta de desenvolver um câncer de mama.

O tratamento mais frequente destas lesões é o acompanhamento com mastologista três vezes ao ano e uso de drogas que diminuem esta incidência (BREASTCANCER,2013).

A Figura 5 ilustra um lóbulo mamário com CLIS

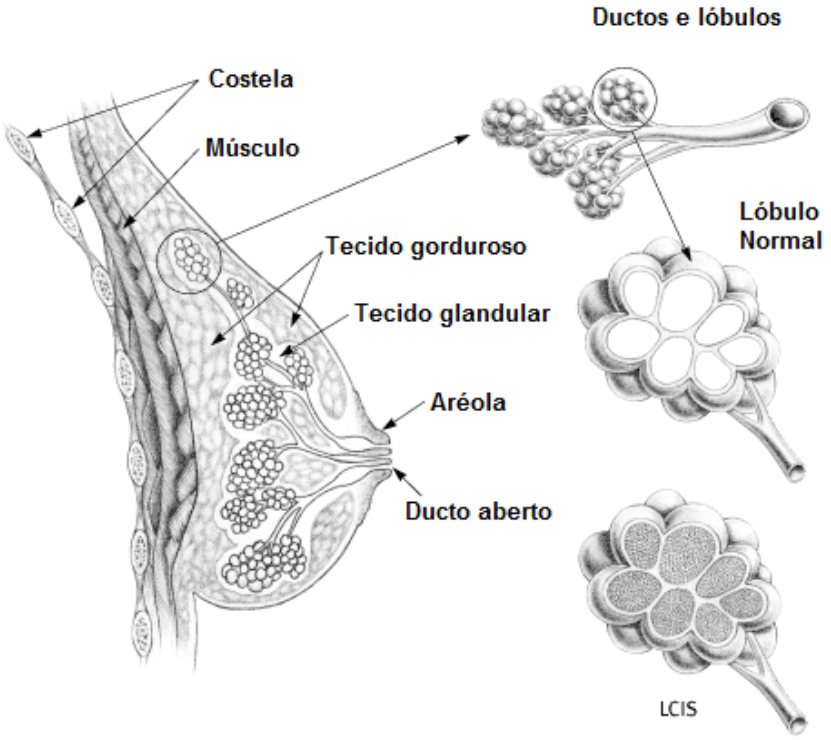


Figura 5 – Ilustração de uma mama com CLIS. Reprodução: (CANCER COUNCIL, 2007)

2.2.1.3. Carcinoma Ductal Invasivo

O Carcinoma Ductal Invasivo (CDI), às vezes chamado de carcinoma ductal invasor, é o tipo mais comum de câncer de mama. Cerca de 80% de todos os cânceres de mama invasivos são carcinomas ductais.

O CDI é um tipo muito comum de câncer de mama. Ele começa a desenvolver nos dutos de leite da mama, mas foge dos tubos do ducto e invade tecidos circundantes. Ao contrário de carcinoma ductal *in situ* (CDIS), que é não invasivo, o CDI tem o potencial de invadir a linfa e o sistema sanguíneo, espalhando células cancerosas para outras partes do seu corpo, originando a metástase (BREASTCANCER, 2013).

2.2.1.4. Carcinoma Lobular Invasivo

O Carcinoma lobular invasivo (CLI) é um tipo de câncer de mama, que começa nos lóbulos da mama, onde o leite é produzido. As células cancerosas que se situam

dentro dos lóbulos infiltram o tecido vizinho fora dos lóbulos. Tal como acontece com carcinoma ductal invasivo (CDI), CLI tem o potencial de disseminação para outras partes do corpo, ocasionando metástase.

Aproximadamente 10 a 15% dos casos de cânceres são diagnosticados como Carcinoma Lobular Invasivo (BREASTCANCER, 2013).

A Figura 6 ilustra mamas com câncer ductal e lobular - ambos invasivos

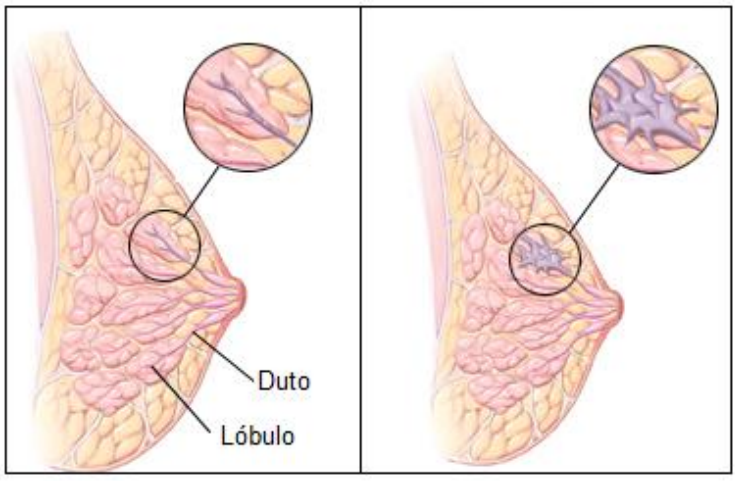


Figura 6 – Ilustração de mamas com câncer: Ductal invasivo (a esquerda) e lobular invasivo (a direita). Reprodução (BREASTCANCER.org, 2013)

2.3. Epidemiologia

Segundo o INCA (2012), a idade continua sendo o principal fator de risco para câncer de mama. As taxas de incidência aumentam rapidamente até os 50 anos de idade, e posteriormente, decrescendo com o aumento da idade. Entretanto, existem outros fatores de risco, que já estão bem estabelecidos, tais como associação ao fator hereditário, que segundo alguns estudos pode ser responsável por 90% dos casos de câncer de mama em todo o mundo, fatores relacionados à vida reprodutiva da mulher (menarca precoce, nuliparidade, primeira gestação tardia, anticoncepcionais orais, menopausa tardia e terapia de reposição hormonal) (TIEZZY, 2009).

Junto com os fatores acima citados, ainda pode-se relacionar a mudança dos padrões globais de vida e saúde, que têm modificado o estilo de vida das mulheres, que nos dias atuais trabalham mais, realizam pouca atividade física, possuem dieta desregulada e agora fazem parte do grupo de risco de desenvolvimento de câncer de mama (TIEZZY, 2009; BRAY, 2004).

Estudos apontam um novo fator de risco, que pode englobar alguns dos fatores já existentes: alta densidade do tecido mamário, que é obtida através da razão entre a quantidade de tecido glandular e o tecido total da mama (STONE et al, 2010; MCCORMACK; SILVA, 2006).

A densidade mamográfica depende de muitos fatores, tais como número de filhos, índice de massa corporal (IMC) e idade, mas o tecido denso é liposubstituído ao longo do tempo, e é aumentado quando existe o uso contínuo de terapia de reposição hormonal pós-menopausa (VERHEUS, 2007; MANDELSON, 2000).

Na literatura recente, observa-se que a densidade de tecido mamário é considerada um fator de risco elevado para o desenvolvimento de um câncer de mama (SHEPHERD, 2011; BOYD, 2007; GIERACH, 2012; YAGHJYAN, 2011).

Mulheres com a densidade mamográfica superior a 75% possuem o risco aumentado de desenvolver câncer de mama entre 4 a 6 vezes, comparado com mulheres que possuem pouco ou nenhum tecido denso na mama (URSIN, 2003; BYRNE, 1995, BOYD, 1995; WOLFE, 1987; BYRNE, 2001; HARVEY, 2004; BOYD, 2005; MCCORMACK, 2006).

Shepherd et al (2011) relata que a densidade mamária é um fator de risco conhecido, entretanto não existe uma definição única de quanto uma mama é densa ou

não, e não existe uma medida padrão para ser usada na prática. Os autores sugerem usar a medida do volume da densidade mamográfica como o melhor preditor de risco.

Gierach et al (2012) estudou a relação densidade mamográfica X mortes por câncer de mama. Segundo os autores, os fatores associados ao desenvolvimento do câncer de mama não são os mesmos fatores que influenciam o risco de morte pela mesma doença. Ainda no artigo, é observado que a densidade mamária é um fator associado ao desenvolvimento, mas não está associado com o óbito por câncer de mama.

O artigo publicado por Yaghjyan (2011) possui os resultados contrários aos encontrados por Gierach (2012), mas com uma ressalva. Segundo os autores, a densidade mamária está associada ao surgimento de câncer, com tumores agressivos com alta mortalidade. Entretanto, a análise foi restrita somente a mulheres no período da pós-menopausa.

Boyd (2007) publicou um artigo no *The New England Jounal of Medicine*, onde foi examinada a associação entre a alta densidade do tecido mamário com todos os fatores de risco conhecidos, tais como: Idade, IMC, idade da menarca, idade do primeiro filho, idade da menopausa, uso de terapia de reposição hormonal e casos de câncer na família. Segundo o autor, mulheres com a densidade do tecido mamário igual ou superior a 75% têm o fator de risco aumentado em 3.5%, se detectado por mamografia. O caso mais grave é quando o câncer é detectado apenas no segundo exame de mamografia (12 meses após o primeiro exame). Neste caso, a razão de probabilidade sobe para 17.8%. A Tabela 2 ilustra uma adaptação dos resultados obtidos no estudo.

Tabela 2 – Densidade Mamográfica e o risco de câncer. Adaptado de (BOYD,2007)

Densidade Mamográfica	Detecção por Mamografia R.P. (95% I.C)	Detecção< 12 meses após o primeiro exame R.P. (95% I.C)
<10%	1.0	1.0
10 a 25%	1.6 (1.2-2.2)	2.1 (0.9-5.2)
25 a 50 %	1.8 (1.3-2.4)	3.6 (1.5-8.7)
50 a 75%	2.0(1.3-2.9)	5.6 (2.1-15.3)
>75%	3.5 (2.0-6.2)	17.8 (4.8-65.9)

2.4. O Diagnóstico Precoce do Câncer de Mama

Os conhecimentos que hoje se têm sobre o câncer de mama são insuficientes para a adoção de programas de prevenção primária, ou seja, medidas que evitem o aparecimento da doença. A maioria dos esforços relacionados ao controle dessa doença está dirigida nas ações de detecção precoce, isto é, na descoberta dos tumores ainda pequenos, com consequente tratamento na fase inicial da doença. Portanto, é muito importante para o diagnóstico da doença, o exame das mamas feito mensalmente pela própria mulher e o realizado pelo médico no decurso de uma consulta de rotina ou não (INCA,2012).

Infelizmente ainda não se dispõe de prevenção com eficácia comprovada e o melhor que se pode fazer é a detecção precoce através da mamografia, que é um método mais acessível.

2.4.1. A Mamografia

A mamografia constitui uma forma particular de radiografia, que trabalha com níveis de tensões e correntes em intervalos específicos, destinada a registrar imagens da mama a fim de diagnosticar a presença ou ausência de estruturas que possam indicar doenças. A Figura 7 ilustra uma mamografia na visão Médio Lateral Oblíqua (MLO)

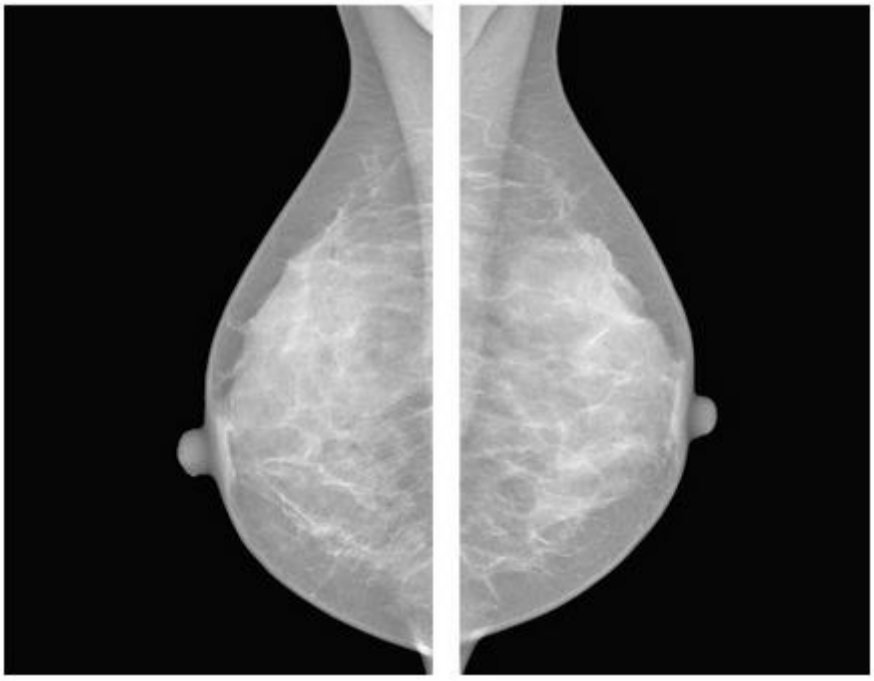


Figura 7– Incidência Médio Lateral das Mamas. Reprodução: (MOREIRA ET AL, 2012)

O reconhecimento de estruturas que possam indicar a presença de câncer se dá através da constatação de uma diferença de contraste entre os diversos tecidos envolvidos.

A gordura, por exemplo, absorve uma menor quantidade de raios-X, aparecendo mais escura no mamograma, enquanto tecidos fibroglandulares apresentam densidade óptica maior e aparecem mais claros (BOYD et al, 1995). Geralmente microcalcificações e massas aparecem em tonalidades mais claras na imagem obtida após a revelação do filme mamográfico, mas essa diferenciação fica prejudicada em imagens de mamas densas, que possuem baixa sensibilidade ao exame (KERLIKOWSKE et al, 1996; GILS, 1998).

Por esse motivo, muitas vezes a descoberta do câncer de mama em mulheres que possuem mamas com tecido denso acontece quando o tumor já apresenta um desenvolvimento avançado, o que dificulta o tratamento da doença.

2.4.1.1. Classificação baseado no tecido

Conforme foi explicitado anteriormente, a alta densidade do tecido mamário pode indicar a presença de tumores malignos. A composição do tecido da mama pode ser um obstáculo na detecção de lesões. Tecido fibroglandular da mama é mais denso que o tecido adiposo, que dificulta a diferenciação entre tecido sadio e tecido lesionado, dificultando o processo de diagnóstico precoce.

Radiologicamente, a densidade do tecido pode esconder tumores, que aumenta a dificuldade de detecção do câncer, aumenta as taxas de reconvocação para novo exame das pacientes, reduz a especificidade da mamografia e compromete a eficácia do exame em mulheres que possuem mamas densas (PERSSON et al, 1997; KERLIKOWSKE et al, 1996;YANKASKAS et al, 2001;TABAR et al, 1995).

Wolfe (1976) foi um dos primeiros pesquisadores a apresentar uma relação entre diferentes densidades de tecido mamário e a probabilidade de desenvolvimento do câncer de mama.

Atualmente, médicos e radiologistas classificam o tecido mamário de acordo com o sistema BI-RADS, desenvolvido pela *American College of Radiology*-ACR (ACR, 2003).

O BI-RADS também é utilizado para categorização do laudo mamográfico, atribuindo desde a categoria "zero" para exames inconclusivos, e "seis" para tumores conhecidos. Nesta tese, será estudada apenas a classificação BI-RADS relacionada ao padrão de tecido mamário.

Segundo o BI-RADS, os padrões mamográficos são divididos em quatro tipos:

- BI-RADS I: Mamas predominantemente adiposas, contendo cerca de até 25% do componente fibroglandular;

- BI-RADS II: Mamas parcialmente adiposas, com densidade de tecido fibroglandular ocupando de 26 a 50% do volume da mama;
- BI-RADS III: Mamas com padrão denso e heterogêneo, nas quais se observa 51 a 75% de tecido fibroglandular, o que pode dificultar a visualização de eventuais nódulos;
- BI-RADS IV: Mamas muito densas, por apresentarem mais de 75% de tecido fibroglandular, o que pode diminuir a sensibilidade da mamografia.

A classificação acima pode ser útil, principalmente nos casos III e IV, onde um acompanhamento especial deve ser realizado, tais como repetição de mamografias a cada 12 meses, verificação minuciosa do histórico da paciente e associação do laudo mamográfico com fatores de risco apresentados. A Figura 8 ilustra quatro mamogramas, classificados segundo o sistema BI-RADS:

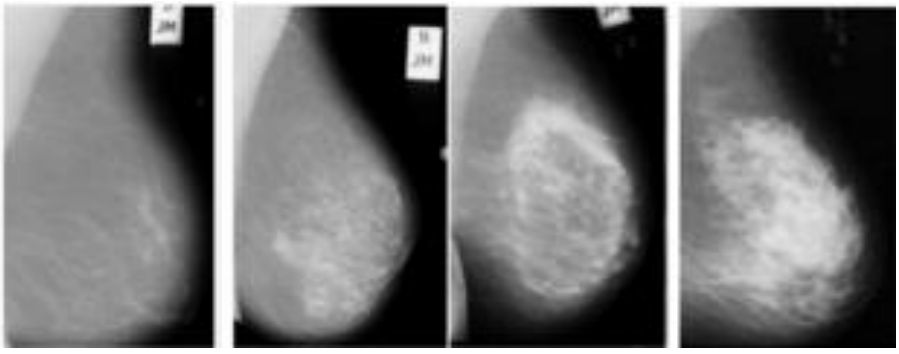


Figura 8 – Mamogramas classificados através da densidade, segundo BI RADS. Da esquerda para a direita: BI-RADS I, BI-RADS II, BI RADS III E BIRADS IV. Reprodução: (MENOTTI; SILVA, 2012)

2.5. Diagnóstico Auxiliado por Computador - CAD

A mamografia é um dos melhores métodos de diagnóstico precoce, sendo a sua interpretação um desafio para o radiologista (MAJID, 2003). Aproximadamente 10 a

30% das lesões mamárias são perdidas, devido a limitações próprias dos observadores humanos, principalmente se a mamografia tiver baixo contraste, como ocorre no caso de mamas com alta densidade de tecido mamário, ocasionando um número elevado de falsos positivos – gerando biópsias desnecessárias – e no pior dos casos, um número elevado de falso-negativos, que ocasiona um retardo no diagnóstico (SUGANTHI et al, 2012; BIRD, 1992).

A necessidade de analisar um grande número de imagens para detectar um pequeno número de casos positivos, o erro de posicionamento ou técnica inadequada de uma mamografia, a localização da lesão fora do campo de visão, características sutis de malignidade associadas ao cansaço ou distração do radiologista, contribuem para interpretações falso-negativas de uma mamografia (MAJID, 2012; PAQUERAULT, 2009).

A dupla leitura em mamografias mostrou-se uma ferramenta significativa, já que foi observada a redução de falso-negativos entre 5 a 15%, melhorando as taxas de detecção (KARSSEMEIJER et al, 2003; SOHNS, 2010), mas apesar de seus benefícios, a utilização nem sempre é possível, devido a limitações financeiras e logísticas de cada instituição.

Com o avanço da tecnologia, do processamento digital de imagens e reconhecimento de padrões, a comunidade cientifica vêm reunindo esforços para desenvolver ferramentas computacionais que possam servir de auxilio ao diagnóstico em imagens. O Diagnóstico Auxiliado por Computador ou Detecção Auxiliada por Computador – CAD é uma ferramenta computacional relativamente recente, que tem sido implementada para prover dupla releitura. Estudos clínicos têm demonstrado que o

CAD aumenta a sensibilidade de detecção do câncer de mama por radiologistas em até 20 a 21% (CALAS et al, 2012).

O Objetivo do CAD não é de diagnóstico, tampouco substituir o radiologista, e sim alertar para áreas específicas, em que uma análise minuciosa irá decidir a necessidade ou não de estudos adicionais.

Os sistemas CAD podem ser utilizados para detecção (segmentação), classificação, ou ambos, se for o caso.

Na detecção, o sistema encontra uma lesão não perceptível ao radiologista. Já na classificação, o radiologista e o sistema vão analisar se região suspeita de conter tecido lesionado possui algum tipo de anormalidade ou não. A forma interessante de CAD seria a combinação de detecção (segmentação) e classificação, pois assim, o sistema iria encontrar uma região suspeita de lesão, não perceptível pelo radiologista, e posteriormente iria classificá-la, em normal ou anormal. (CALAS et al, 2012).

2.5.1. Desempenho de Sistemas CAD sobre Mamas Densas

Embora existam poucos ou quase nenhum sistema CAD específico para mamas densas, trabalhos relacionados na literatura relatam seu desempenho, quando utilizados com mamas gordurosas e mamas densas. Segundo Yang (2007), a prevalência de marcadores falso-positivos aumenta com o aumento da densidade mamária. Obenauer (2006) observou a tendência de que a densidade mamária pode afetar a detecção por CAD. Ho e Lam (2003) mostraram redução da sensibilidade estatística do CAD com o aumento da densidade mamária. A sensibilidade do CAD foi de 93,3%, com especificidade de 1,3 falso-positivo por imagem nos casos de mamas adiposas, entretanto, reduziu para 64,3% para mamas muito densas.

Alguns trabalhos, como o de Brem (2005), não foi encontrado nenhuma diferença significativa da detecção de câncer entre as mamas densas e não densas. Entretanto, a acurácia total do método desenvolvido foi relativamente baixa, em torno de 89%, fato que pode ser associado a um classificador insatisfatório. Mesmo com tais resultados, os autores sugerem que sistemas CAD podem ser particularmente vantajosos em pacientes com mamas densas, nas quais a mamografia é mais desafiadora.

Em um trabalho mais recente, Pinker et al (2010) relatou que dos 200 casos de câncer estudados, o sistema proposto marcou corretamente 79%, e evidenciou que a baixa sensibilidade está associada à densidade do tecido.

Baseado em tais resultados, observa-se a necessidade de um sistema específico para mamas densas, pois o grande problema dos CAD atuais está justamente em obter uma acurácia satisfatória neste tipo de mama.

2.6. Processamento Digital de Imagens

Técnicas de processamento de imagens digitais surgiram, principalmente, pela necessidade de melhorar a qualidade das imagens e fornecer outros subsídios que facilitem a interpretação humana. Ao longo das duas últimas décadas, a área de processamento digital de imagens experimentou um rápido crescimento, expandindo a cada dia o domínio de aplicações e soluções possíveis. Podemos citar como exemplo, os exames de diagnóstico por imagens, que hoje em dia são ferramentas indispensáveis em hospitais e clínicas.

Aplicado ao presente trabalho, o processamento de imagens é dividido em cinco itens, conforme descritos abaixo:

1. Aquisição de Imagens:

   Na aquisição utiliza-se algum mecanismo para gerar as imagens que se deseja processar. No caso do trabalho proposto foi utilizada a base de mamografias digital disponível na base de dados radiográficas MIAS - *Mammographic Institute Analisys Society* (SUCKLING et al, 1994).

2. Pré-Processamento:

   O pré-processamento tem a finalidade de aumentar a qualidade da imagem, eliminar objetos indesejáveis, tornando mais fácil a sua identificação e interpretação.

   Neste trabalho, foram retiradas as etiquetas de identificação, falhas na digitalização e o músculo peitoral de cada mamografia utilizando operadores morfológicos de erosão e dilatação e crescimento de regiões.

3. Segmentação:

   Permite o isolamento do objeto de estudo, onde os seus resultados são cruciais na determinação de sucesso ou falha na análise da imagem. A segmentação deve focar em isolar somente as Regiões Suspeitas de conterem uma lesão, aqui chamadas de ROS – *Region on Suspicious*. É uma fase delicada neste trabalho, pois está relacionada com características da imagem que são difíceis de traduzir para a máquina. A dificuldade consiste em encontrar medidas consistentes que possam levar a máquina a decidir corretamente a que grupo cada *pixel* pertence. Para esse trabalho utilizamos Equalização Adaptativa de Histograma com Contraste Local somado ao Alargamento de Contraste (Seção 2.6.2 e Seção 2.6.3).

4. Extração de Características

   Possui a finalidade de extrair da região segmentada um conjunto descritivo de características mensuráveis. Estas características devem variar de acordo com a região segmentada. Por exemplo, uma região que contenha algum tipo anormalidade - tal como um tumor benigno ou maligno - deve possuir características mensuráveis bem distintas, se comparada à região que possui apenas tecido normal, sem qualquer tipo de displasia ou neoplasia. Deve-se utilizar medidas que resultem em informações importantes para discriminação entre classes distintas. O conjunto dessas medidas constitui um vetor de características que definem um padrão calculado para aquela determinada região segmentada. Neste trabalho, as regiões de interesse foram descritas através da Análise de Componentes Independentes (seção 2.6.3).

5. Classificação

   Busca através do vetor de características obtido na etapa de extração de características, classificar o objeto em algum grupo determinado previamente, no caso desta pesquisa, normal ou anormal. Nesta tese, utilizou-se uma técnica de aprendizado supervisionado Máquinas de Vetores de Suporte – SVM (seção 2.6.5) visando reconhecer os padrões existentes nas características das regiões de suspeitas encontradas e classificá-las em normal ou anormal.

As próximas subseções apresentam as principais técnicas utilizadas no desenvolvimento deste trabalho.

2.6.1. Equalização Adaptativa de Histograma com Limitação de Contraste

O histograma de uma imagem é a representação gráfica de um conjunto de números, indicando o percentual de *pixels* naquela imagem que apresentam um determinado nível de cinza. Estes valores são normalmente representados por um gráfico, que fornece para cada nível de cinza o número (ou o percentual) de *pixels* correspondentes na imagem. Através da visualização do histograma de uma imagem obtém-se uma indicação de sua qualidade quanto ao nível de contraste enquanto ao seu brilho médio (se a imagem é predominantemente clara ou escura) (MARQUES, VIEIRA, 1999).

Cada elemento deste conjunto é calculado por:

$$P(r) = \frac{h_j}{h} \quad , \tag{2.1}$$

sendo:

- j = 0, 1, 2, ..., L-1 , onde L é o número de níveis de cinza da imagem
- h = Número total de *pixels* da imagem
- $p_{(r)}$= Probabilidade do j-ésimo nível de cinza
- h_j= Número de *pixels* cujo nível de cinza corresponde a j

A equalização de histograma é uma técnica a partir da qual se procura redistribuir os valores de tons de cinza dos *pixels* em uma imagem, de modo a obter um histograma uniforme, no qual o percentual de *pixels* de qualquer nível de cinza é praticamente o mesmo. Para tanto, utiliza-se uma função auxiliar, denominada função de transformação. A forma mais usual de se equalizar um histograma é utilizar a função de distribuição cumulativa (CDF) da distribuição de probabilidades original, que pode ser expressa por: (GONZALEZ, WOODS, 2007).

$$G_j = T(r_j) = \sum_{l=0}^{j} \frac{h_j}{h} = \sum_{l=0}^{k} p_r(r_l) \quad , \tag{2.2}$$

sendo:

- $0 \leq r_k \leq 1$
- k = 0, 1, 2, ..., L-1 , onde L é o número de níveis de cinza da imagem

Uma variação da técnica tradicional de Equalização de Histograma é a Equalização Adaptativa de Histograma, que divide a imagem em pequenas regiões, chamadas de "regiões contextuais". Dessa forma, em vez de trabalhar com a imagem inteira, tal como a equalização de histograma tradicional, a equalização adaptativa modifica os *pixels* baseado em uma pequena região contextual, com poucos vizinhos. Os vizinhos das fronteiras de cada região são combinados utilizando interpolação bilinear, a fim de evitar bordas, artificialmente induzidas (PIZER, AMBURN, 1987).

O resultado da Equalização Adaptativa é uma imagem com contraste local ampliado, apresentando mais detalhes. A grande desvantagem é que o ruído/saturação gerado em regiões de baixo contraste é ampliado junto com a imagem.

Uma alternativa para resolver a desvantagem da Equalização Adaptativa é limitar o contraste, modificando o aclive da função de transformação do histograma e redistribuir os *pixels* novamente, conforme ilustra a Figura 9 (PIZER, AMBURN, 1987; SUNDARAMI et al. 2011).

O algoritmo mais usual que realiza a limitação do contraste em histogramas pode ser descrito pela fórmula abaixo: (RAI et al. 2012)

$$g = g_{min} - \left(\frac{1}{\alpha}\right) * ln\left[1 - \arg(G_j)\right] \tag{2.3}$$

Sendo:

- g = Valor do *pixel* atualizado

- g_{min} = Valor do menor *pixel*
- G_j = Função de Distribuição cumulativa
- α = *Clip limit*

A Figura 9 ilustra a nova distribuição dos *pixels* de um histograma gerado através de equalização adaptativa após a utilização da técnica de limitação de contraste, também chamada de *Contrast Limited Adaptive Histogram Equalization* – CLAHE. Pode- se notar que um novo valor de *clip* é gerado, pois os *pixels* situados acima do *clip* são remapeados ao longo do eixo de intensidade L-1.

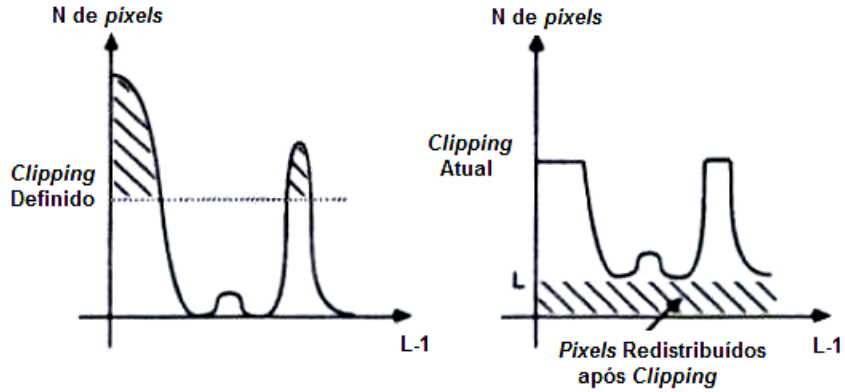


Figura 9 – Ilustração de redistribuição de *pixels* após aplicação de técnicas de limitação de contraste. O gráfico a esquerda ilustra um histograma equalizado, apresentando um limiar (aqui chamado de clipping). O gráfico a direita ilustra a redistribuição dos *pixels* ao longo do eixo de intensidade (L-1), e apresenta o *clipping* atual, modificado devido a nova distribuição. Reprodução: (PIZER,1987)

2.6.2. Alargamento de Contraste

Alargamento de Contraste é uma das mais simples transformações não- lineares definidas por partes. Imagens de baixo contraste resultam em baixa riqueza de detalhes, sendo de difícil visualização e interpretação. O Alargamento de Contraste é uma técnica que expande uma determinada faixa de níveis de intensidade de modo a incluir todo o intervalo, sendo que a expansão (alargamento) ocorre somente no intervalo determinado pelos parâmetros da função de transformação linear por partes.

A Figura 10 mostra uma transformação típica utilizada para o alargamento de contraste. As posições dos pontos (r_1 e s_1) e (r_2 e s_2) controlam o formato da função de transformação.

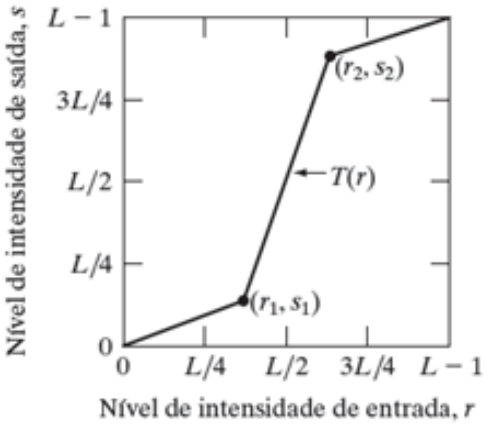


Figura 10 – Função de transformação não linear para alargamento de contraste

2.6.3. Análise de Componentes Independentes

A análise de componentes independentes, do inglês *Independent Component Analysis* (ICA) é um método computacional desenvolvido inicialmente, para resolver problemas de Separação Cega de Fontes, do inglês *Blind Source Separation* (BSS) (HYVÄRINEN, KARHUNEN e OJA, 2001).

Uma aplicação de ICA bastante comum é a extração de características, que é aplicada em várias situações que envolvam processamento de sinais e imagens (CAMPOS *et al*, 2007; COSTA *et al*, 2011; ARONS, 1990; VIGARIO, 1997). Em processamento de imagem, as componentes podem fornecer uma representação para uma imagem. Tal representação permite executar tarefas como compressão ou reconhecimento de padrões (HYVÄRINEN, KARHUNEN e OJA, 2001).

2.6.3.1.Definições

Sejam dadas observações de *n* sinais, modelados como combinações lineares de *n*

funções bases:

$$x_n = a_1 s_1 + a_2 s_2 + \cdots + a_n s_n \tag{2.3}$$

Sendo

- x_n= Sinal aleatório
- s_n = Componente independente aleatório
- a_n= Coeficiente de mistura

Utilizando notação matricial, podemos reescrever esta equação da seguinte forma:

$$\mathbf{X=A.s} \tag{2.4}$$

O modelo apresentado na equação (2.4) é chamado de Análise de Componentes Independentes - ICA (HYVÄRINEN, OJA, 2000), que descreve como os dados são gerados a partir do processo de mistura com as componentes independentes.

O objetivo deste modelo é permitir que se estime a matriz de mistura **A**, bem como a matriz de componentes independentes **S**, somente observando **X**.

A estimação das componentes é baseada nas seguintes condições:

- As componentes independentes são estatisticamente independentes;
- As componentes possuem distribuição não-gaussiana.

O modelo de ICA apresenta, no entanto, algumas ambiguidades no que diz respeito às componentes independentes:

- Não se pode determinar suas variâncias;
- Não se pode determinar sua ordem.

Tais ambiguidades se devem ao fato de **A** e **S** serem desconhecidas. Como

consequência, não é possível determinar as energias ou as amplitudes dos sinais, nem tão pouco os sinais ou a ordem de s_n (HYVÄRINEN *et al*, 2001).

Generalizando o modelo acima para processamento de imagens, uma imagem será considerada uma mistura de imagens-base, ou seja, uma combinação linear de algumas imagens-base com os coeficientes a_n, considerados como características da imagem.

A Figura 11 ilustra a representação de uma imagem como combinação linear de suas imagens-base.

Figura 11– Imagem como uma mistura de imagens mutua e estatisticamente independentes entre si

2.6.3.2. Descorrelação e Independência

Duas variáveis são consideradas independentes quando o valor de uma não fornece informação acerca do valor da outra. Consideremos duas variáveis x_1 e x_2 . Estas variáveis são ditas independentes se, e somente se, x_1 não fornece nenhuma informação de x_2 e vice-versa. Matematicamente,

$$p(x_1, x_2) = p(x_1).p(x_2) \tag{2.5}$$

Ou usando outros termos, pode-se dizer que a probabilidade conjunta de x_1 e x_2 é igual ao produto das densidades marginais $p(x_1)$ e $p(x_2)$.

Duas variáveis x_1 e x_2 são descorrelacionadas se a sua covariância for igual a zero, ou seja:

$$cov_{x1,x2} = E[(x_1 - \mu_1)]E[(x_2 - \mu_2)] = 0 \tag{2.6}$$

Sendo μ_1 e μ_2 as médias das variáveis x_1 e x_2, respectivamente.

2.6.3.3. Estimação das Componentes Independentes

A estimação das componentes independentes s_n pode ser obtida através da matriz de mistura **A**, da seguinte forma:

$$\mathbf{S} = \mathbf{A^{-1}X} \tag{2.7}$$

Sendo a matriz **A** desconhecida, a ideia principal da análise de componentes independentes consiste em considerar que os sinais observáveis x_n estão relacionados com os sinais originais através de uma transformação linear. Assim, os sinais originais podem ser obtidos a partir de uma transformação inversa. Supondo, dessa forma, uma combinação linear de x_i, de modo que:

$$y = b^T X \, . \tag{2.8}$$

Sendo **X=AS**, pode-se escrever:

$$y = b^T \mathbf{AS} \ . \tag{2.9}$$

Onde b deve ser determinado. A partir da equação (2.9), é possível observar que *y* é uma combinação linear se s_i, com coeficientes dados por $q=b^T\mathbf{A}$. Sendo assim, obtêm-se:

$$y=q^T\mathbf{S} \tag{2.10}$$

Se *b* corresponde a uma das linhas da inversa de **A**, então *y* será uma das componentes independente, e neste caso, apenas um dos elementos de *q* será igual a um, e todos os outros serão iguais a zero. No entanto, sendo **X** conhecido, *b* não pode ser determinado exatamente, porém pode-se estimar seu valor.

Uma forma de determinar *b* é variar os coeficientes em *q* e verificar como a distribuição de $y=q^T\mathbf{S}$ muda. Como pelo Teorema do Limite Central (PAPOULIS, PILLAI, 2002), a soma de variáveis aleatórias independentes de média finita e variância limitada é aproximadamente normal, desde que o número de termos da soma seja suficientemente grande. Assim, apenas um elemento q_i de *q* é diferente de zero

(HYVÄRINEN *et al*, 2001). Como, na prática, os valores de *q* são desconhecidos, e através da equação 2.8 e da equação (2.10) tem-se que:

$$b^T\mathbf{X}=q^T\mathbf{S} \tag{2.11}$$

Pode-se variar b e observar a distribuição de $b^T\mathbf{X}$

Dessa forma, pode-se tomar como *b* um vetor que maximiza a não-gaussianidade de $b^T\mathbf{X}$, sendo $q=\mathbf{A^TS}$, contendo apenas uma de suas componentes diferente de zero. Isso significa que *y* na equação 3.8 é igual a uma das componentes independentes, e a maximização da não-gaussianidade de $b^T\mathbf{X}$, permite encontrar uma das componentes.

2.6.3.4. Negentropia como Medida de Não Gaussianidade

A entropia de uma variável aleatória está relacionada com a quantidade de informação que essa variável possui. Sendo *y* um vetor aleatório com função densidade de probabilidade *f(y)*, a sua entropia diferencial é dada por:

$$H(y) = -\int f(y) log f(y) dy \tag{2.12}$$

Sabendo-se que uma variável gaussiana tem a maior entropia dentre todas as variáveis aleatórias de igual variância (HYVÄRINEN *et al*, 2001; PAPOULIS,PILLAI, 2002), tem-se que uma versão modificada da entropia diferencial pode ser usada como medida de não-gaussianidade. Tal medida é denominada negentropia, definida por:

$$\mathrm{H(y)} = \mathrm{H}\left(\mathrm{y_{gauss}}\right) - \mathrm{H(y)} \tag{2.13}$$

sendo $\mathrm{y_{gauss}}$uma variável aleatória de mesma matriz de covariância que y. A negentropia é sempre não negativa, e pode assumir zero se, e somente se, y tem distribuição gaussiana e é invariante para transformações lineares inversíveis.

Apesar de permitir que se possa medir não-gaussianidade, a negentropia é de difícil estimação, sendo necessária sua estimação por aproximações através de momentos de alta ordem. Assim,

$$J(y) \approx \frac{1}{12}E\{y^3\}^2 + \frac{1}{48}kurt(y)^2 \quad (2.14)$$

Sendo $kurt(y)$ a *kurtosis* de y, definida como o momento de quarta ordem da variável aleatória y, definida por

$$kurt(y) = E\{y^4\} - 3(E\{y^2\})^2 \quad (2.15)$$

2.6.4. Seleção de Características Mais Significantes

Identificar as características mais importantes dentre um vetor de características observado, é uma das tarefas mais críticas encontradas em sistemas de reconhecimento de padrões. Tal tarefa é considerada de essencial importância para diminuir o erro de classificação e o custo computacional **(**PENG *et al*, 2005).

As características irrelevantes podem ser removidas sem comprometer o resultado da classificação, pois neste contexto, são consideradas redundantes, ou seja, implicam na presença de outra característica com a mesma funcionalidade, e não trás nenhuma informação nova ao vetor de características.

Seja *v* um vetor de dados, com n amostras e m características $v = \{v_i, i = 1, \cdots M\}$ e seja *c* um vetor de classe (rótulo).

O problema de seleção de características (do inglês *feature selection*) é encontrar do espaço de observação de dimensão *v*, um subespaço de m características que represente "otimamente" *c*.

Dada a condição de encontrar a "caracterização ótima", o algoritmo deve buscar a melhor forma de encontrar este subespaço (de caracterização ótima). As duas formas mais comuns são: classificando-as por algum critério e selecionando as *k* melhores características, a segunda é escolher um subconjunto mínimo dentro do conjunto de características sem afetar a precisão da classificação.

Sendo assim, na seleção de um subconjunto "ótimo" os algoritmos podem automaticamente determinar o número de características, ou o operador pode estipular o tamanho do subespaço de características.

A condição de "caracterização ótima" implica em um erro de classificação mínimo, que requer a máxima dependência estatística entre o subespaço m selecionado, e o vetor de classe *c*. Tal esquema é chamado de Máxima Dependência.

Em termos de informação mútua, a proposta de realizar a seleção de características para encontrar um vetor *v'*, com m características $\{v_i\}$, que conjuntamente elas tenham a maior dependência possível com o vetor de classe *c*, é dada por:

$$maxD(v,c), \quad D = I(\{v_i, i = 1, \cdots, m\}: c) \tag{2.16}$$

Entretanto, a equação acima pode demandar um esforço computacional grande, quando os dados são multivariados, já que para estimar o vetor *v'*, é necessário calcular inúmeras vezes a inversa da matriz de covariância. Devido ao fato exposto, o algoritmo para encontrar a máxima dependência entre variáveis é considerado de grande esforço computacional, ocasionando um custo computacional elevado.

Para diminuir o esforço computacional, pode-se utilizar um critério, baseado em máxima Relevância e Mínima Redundância (mRMR), que maximiza a informação mútua e minimiza a medida de redundância (PENG *et al.*, 2005).

2.6.4.1. Máxima Relevância e Mínima Redundância (mRMR)

Conforme visto nos parágrafos anteriores, o critério de máxima dependência possui custo computacional elevado. Como alternativa, pode-se selecionar características através do critério de Máxima Relevância, que encontra características satisfazendo a

equação abaixo que aproxima $D(v,c)$ da equação 2.16, com o valor médio de todos os valores de informação entre as características individuais x_i e o vetor de classe **c**:

$$maxD(v,c), \quad D = \frac{1}{|v|}\sum_{v_i \epsilon\, d} I(v_i;c) \tag{2.17}$$

É provável que as características selecionadas de acordo com o critério acima descrito tenham muita redundância, ou seja, a dependência entre estas características pode ser grande. Para resolver tal problema, aplica-se em conjunto, a condição de Mínima Redundância, que seleciona mutuamente apenas as mutuamente características exclusivas: (DING, PENG, 2003)

$$minR(v), \quad R = \frac{1}{|v|^2}\sum_{v_i,v_j \in v} I(v_i, v_j) \tag{2.18}$$

Os critérios descritos em 2.17 e 2.18 são chamados conjuntamente de Máxima-Relevância-Mínima-Redundância (mRMR) (PENG, 2005)

Pode-se definir o operador $\phi(D,R)$ para combinar D e R, para em seguida otimizá-los simultaneamente:

$$max\phi(D,R), \phi = D - R \tag{2.19}$$

2.6.5. Máquina de Vetor de Suporte

A Máquina de Vetores de Suporte, do inglês *Support Vector Machine* - SVM é um método de aprendizagem supervisionada capaz de classificar a partir de *n* indivíduos observados pertencentes a diversos subgrupos, a que classe um indivíduo que deve ser classificado pertence (BISHOP, 2006).

SVM vêm sendo considerado um dos melhores classificadores não lineares, e é bastante utilizado em trabalhos que envolvam detecção de câncer em mamografias digitais (COSTA *et al*, 2011; MARTINS *et al*, 2009).

A ideia da SVM é construir um hiperplano como superfície de decisão, de tal forma que a margem de separação entre as classes seja máxima possível. O objetivo do treinamento através de SVM é a obtenção de hiperplanos que dividam as amostras de tal maneira que sejam otimizados os limites de generalização.

As SVM são consideradas sistemas de aprendizagem que utilizam um espaço de hipóteses de funções lineares em um espaço de muitas dimensões. Em casos em que o conjunto de amostras é composto por duas classes separáveis, um classificador SVM 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. Mesmo quando as duas classes não são separáveis, a SVM é capaz de encontrar um hiperplano através do uso de conceitos pertencentes á teoria da otimização (VAPNIK, 1998).

A Figura 12 ilustra hiperplanos de separação entre duas classes linearmente separáveis. O hiperplano ótimo (linha tracejada) separa as duas classes, com o auxilio dos vetores de suporte (linhas continuas) e mantém a maior distância possível com relação aos pontos da amostra.

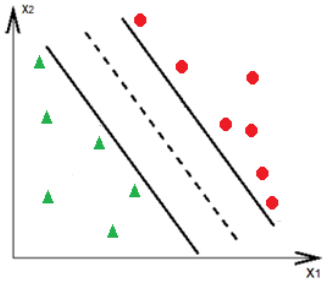


Figura 12 – Separação de duas classes, com o auxilio de vetores de suporte.

2.6.5.1.Definições

Para simplificar, será considerado o caso de classificação utilizando duas classes usando um modelo linear descrito por:

$$y(z) = w^T z + \epsilon = 0 \tag{2.20}$$

sendo w é um vetor de pesos ajustados, ϵ é um viés e z é um vetor de treinamento de características, com seus respectivos rótulos $y_i \in Y$, em que $Y = \{-1, +1\}$. O modelo definido na equação (2.20) define um hiperplano ótimo, que classifica todos os vetores de treinamento, e z é dito linearmente separável se é possível separar os dados das classes –1 e +1 por este hiperplano (SMOLA, SCHÖLKOPF, 2002). A Figura 13 ilustra um hiperplano ótimo para padrões linearmente separáveis.

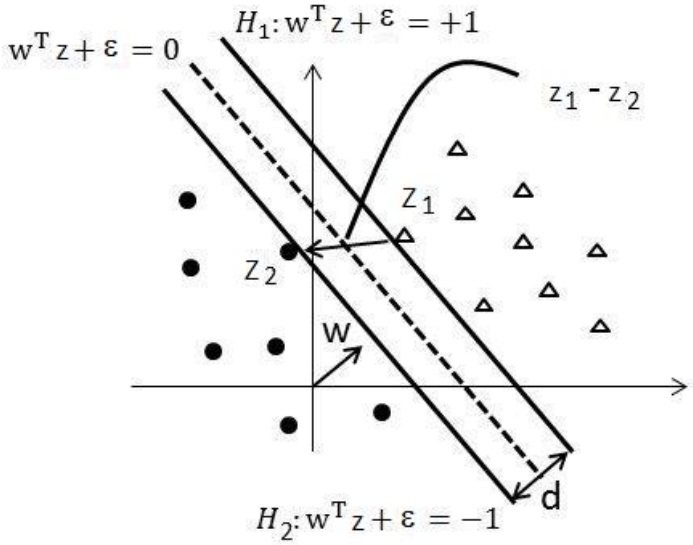


Figura 13 - Hiperplano ótimo, com dois vetores de suporte H_1 e H_2

Baseado no modelo,

$$w^T z + \epsilon \geq 0, \text{ para y(z)=+1} \tag{2.21}$$

$$w^T z + \epsilon < 0, \text{ para y(z)=-1} \tag{2.22}$$

Sendo x_1 um ponto pertencente ao hiperplano $H_1 = w^T z + \epsilon = +1$, e x_2 um ponto pertencente ao hiperplano $H_2 = w^T z + \epsilon = -1$, conforme ilustrado na Figura 14.

Ao se projetar $z_1 - z_2$ na direção de $\boldsymbol{w}$, perpendicular ao hiperplano ótimo, é possível obter a distância entre os hiperplanos H_1 e H_2, dada pela equação abaixo:

$$(z_1 - z_2)\left(\frac{w}{\|w\|} \cdot \frac{(z_1 - z_2)}{\|z_1 - z_2\|}\right) \tag{2.23}$$

De acordo com as equações 2.21 e 2.22, tem-se que $H_1 = w^T z + \epsilon = +1$ e $H_2 = w^T z + \epsilon = -1$. A diferença entre as duas equações dão como resultado $w.(z_1 - z_2) = 2$. Substituindo o resultado encontrado na equação 3.23, tem-se:

$$\frac{2(z_1 - z_2)}{\|w\|.\|z_1 - z_2\|} \tag{2.24}$$

Tomando-se a norma da equação 2.24 acima, tem-se:

$$\frac{2}{\|w\|} \tag{2.25}$$

Esta é a distância d, ilustrada na Figura 14, entre os hiperplanos H_1 e H_2, paralelos ao hiperplano ótimo separado. Minimizando $\|w\|$, pode-se minimizar a margem de separação dos dados em relação ao $w^T z + \epsilon = 0$. Assim, tem-se o problema de otimização descrito por:

$$min_{w,\epsilon} \frac{1}{2} \|w\|^2 \tag{2.26}$$

Com a restrição $y_i(w^T z_i + \epsilon) - 1 \geq 0, \forall_i = 1,2,\ldots,n$, que é imposta para assegurar que não haja dados de treinamento entre as margens de separação das classes. Este é um problema de otimização quadrático, cuja função objetiva é convexa e os pontos que satisfazem as restrições formam um conjunto convexo, logo possui um único mínimo global.

Para solucionar tal problema, aplica-se um operador Lagrangiano, capaz de englobar as restrições às funções objetivo, associadas aos multiplicadores de Lagrange, conforme a equação abaixo:

$$K(w, \epsilon, \rho) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} \rho_1(y_i(w^T z_i + \epsilon) - 1) \tag{2.28}$$

Onde o operador Lagrangiano deve ser minimizado, implicando na minimização das variáveis α_i e na minimização de w e ϵ.

Igualando as derivadas de K em relação a ϵ, e a w a zero, obtemos as condições abaixo:

$$w = \sum_{i=1}^{n} \rho_i y_i z_i \tag{2.29}$$

$$\sum_{i=1}^{n} \rho_i y_i = 0 \tag{2.30}$$

Substituindo as equações (2.29) e (2.30) na equação (2.28), tem-se o seguinte problema de otimização exposto nas equações abaixo:

$$\max_{\rho} \sum_{i=1}^{n} \rho_1 - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_i \rho_j y_i y_j \left(z_i . z_j\right) \tag{2.31}$$

$$\text{Sujeito a } \begin{cases} \alpha_i \geq 0, \forall_i = 1, \dots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases} \tag{2.32}$$

Sendo que $\left(z_i . z_j\right)$ corresponde ao produto interno entre $z_i \; e \; z_j$

Como se trata de padrões não separáveis linearmente, não é possível construir um hiperplano ótimo de separação sem encontrar erros de classificação (HAYKIN, 1999). Para o caso de pontos de dados não separáveis, é introduzido um conjunto de variáveis escalares não negativas, ε_i, na definição do hiperplano de separação:

$$u_i(w^T z + \rho) \geq 1 - \varepsilon_i, \;\; i = 1, 2, \cdots, n \tag{2.33}$$

As variáveis $\varepsilon_i$são chamadas variáveis de folga e medem o desvio de um ponto de dado na condição ideal de separabilidade de padrões. Dessa forma, o problema agora é:

$$\min_{w,\varepsilon} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \varepsilon_i \tag{2.34}$$

Sujeito a: $y_i(w . z_i + \epsilon) \geq 1 - \varepsilon_i, \;\; \varepsilon_i \geq 0, \;\; \forall_i = 1, \cdots, n$ (2.35)

Novamente, trata-se de um problema de otimização quadrático, com as restrições lineares dadas pela equação 3.35. Aplicando o operador Lagrangiano:

$$\max_{\rho} \sum_{i=1}^{n} \rho_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_i \rho_j y_i y_j \left(z_i . z_j\right) \tag{2.36}$$

$$\text{Sujeito a } \begin{cases} 0 \leq \rho_i \geq C, \forall_i = 1, \ldots, n \\ \sum_{i=1}^{n} \rho_i y_i = 0 \end{cases} \tag{2.37}$$

Sendo *C* um parâmetro positivo especificado pelo usuário.

Os problemas não lineares de classificação são resolvidos mapeando o conjunto de treinamento, saindo de seu espaço de entrada para um novo espaço com maior dimensão, denominado espaço de características. Seja $\theta(z)$, uma função que mapeia o espaço de entrada sobre o espaço de características. Esta função é usada para mapear $z_i \; e \; z_j$ para o espaço de características, antes da realização do produto interno entre eles.

$$k(z_i, y_i) = \theta(z_i).\theta(z_j) \tag{2.38}$$

modificando desta forma, o problema de maximização proposto na equação 2.38, para:

$$\max_{\rho} \sum_{i=1}^{n} \rho_1 - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_i \rho_j y_i y_j k(z_i . z_j) \tag{2.39}$$

Com as restrições da equação 2.39. A equação (2.38) é chamada de função de núcleo (*kernel).*

Uma função de núcleo bastante utilizada em SVM é a Função de Base Radial, do inglês *Radial Basis Function* – RBF, e é descrita por:

$$k(z_i . z_j) = e^{-\gamma \|z_i - z_j\|^2} \tag{2.40}$$

Sendo γ o parâmetro, que é definido a priori, pelo usuário.

3. OBJETIVOS

- Propor um sistema de Diagnóstico Auxiliado por Computador específico para mamas densas, que auxilie na detecção precoce do câncer de mama, aumentando as chances de cura da paciente e diminuindo os casos de mortalidade da doença;
- Avaliar o sistema proposto através da sensibilidade, especificidade, acurácia e curva ROC.

4. Materiais e Métodos

O fluxograma do método proposto é mostrado na Figura 14. Consiste na aquisição da imagem, no pré-processamento para retirar artefatos e o músculo peitoral, a segmentação das regiões suspeitas de lesão, a extração de características das regiões suspeitas, a seleção das características mais significantes e, finalmente, a classificação das regiões suspeitas, como normais ou anormais.

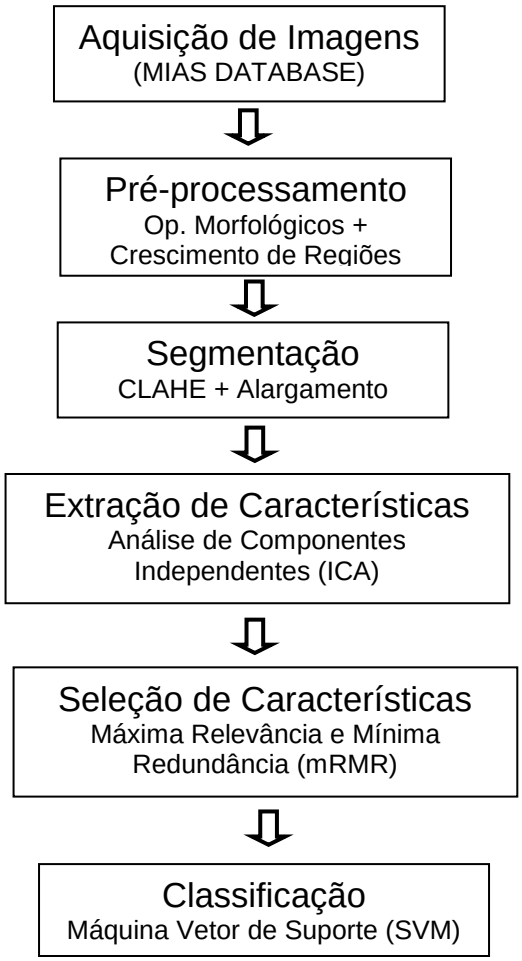


Figura 14–Etapas Aplicadas na Metodologia Proposta

O restante do capítulo descreve cada uma dessas etapas em detalhes, aborda a base de dados utilizada nos testes e as técnicas utilizadas para o desenvolvimento da metodologia proposta.

4.1. MIAS Database

A base de dados utilizada neste trabalho foi cedida pelo *Mammographic Institute Analisys Society.*(SUCKLING *et al*, 1994) e é composta por 322 mamogramas (mamas esquerda e direita) de 161 pacientes, sendo 53 com diagnóstico maligno, 69 com diagnóstico benigno e 206 de diagnóstico normal. As mamografias têm um tamanho de 1024x1024 *pixels* e resolução de 200 *micron*. As anormalidades são classificadas de acordo com a classe encontrada (calcificação, massa circunscrita, distorções arquiteturais, assimetrias, e outras massas, sem formas definidas).

Esta base de dados contém também um arquivo, explicando detalhadamente cada mamografia, como por exemplo, a classificação do tecido da mama, em:

- Gorduroso (F),
- Gorduroso-Glandular (G), que é considerado uma mama em processo de liposubstituição, que apresenta tanto tecido denso como tecido gorduroso,
- Denso (D)

A base de dados também possui informações da classe da anormalidade encontrada, as coordenadas xy do centro da anormalidade e o raio aproximado, em *pixels*.

Das 322 mamografias, existem 122 diagnósticos anormais (maligno ou benigno), sendo 44 encontradas em mamografias com tecido gorduroso (F), 38 em tecido Gorduroso-Glandular (G) e 40 em tecido denso (D), conforme informações contidas na base de dados MIAS.

Neste trabalho, foram utilizadas todas as mamografias com diagnóstico anormal (maligno ou benigno) que foram encontradas em tecidos densos (D) ou gorduroso-glandular.

4.2. Pré-processamento

Imagens de mamografia apresentam elementos típicos do exame que podem interferir no resultado da segmentação, tais como, etiquetas de identificação, ruídos, artefatos, etc., que são gerados no momento da aquisição da imagem. O maior objetivo da etapa de pré-processamento é remover tais elementos típicos e melhorar a discriminação visual das estruturas internas da mama.

A Figura 15 ilustra uma mamografia com elementos típicos que podem alterar o resultado da segmentação.

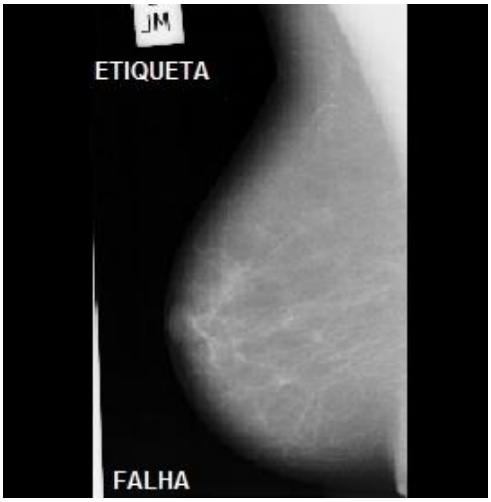


Figura 15 – Mamografia com etiqueta e falhas na digitalização, onde se observa uma etiqueta (acima), e uma falha (abaixo). Adaptado de (SUCKLING et al, 1994)

4.2.1. Extração de etiquetas e falhas na digitalização

A etapa de extração de etiquetas e falhas na digitalização foi realizada utilizando uma função *thresholding*,definido por:

$$g(x,y) = \begin{cases} 1, & se\ f(x,y) \geq t \\ 0, & se\ f(x,y) < t \end{cases}$$

Sendo $f(x,y)$ o valor do *pixel* da imagem original, e $g(x,y)$ a imagem modificada, com *pixels* entre 0 (zero) e 1 (hum) (GONZALEZ, WOODS, 2007)

Dessa forma, com t=15 foi gerada uma máscara binária, com valores de pixels "0" e "1".

Após a elaboração da máscara, é necessária a utilização de operadores morfológicos clássicos para remover todas as etiquetas e falhas em digitalização (GONZALEZ, WOODS, 2007).

Neste trabalho, foram utilizados operadores morfológicos de abertura e fechamento. O operador morfológico de abertura é descrito por:

$$I \circ B = (I \ominus B) \oplus B \tag{4.1}$$

O operador morfológico de fechamento é descrito por:

$$I \cdot B = (I \oplus B) \ominus B \tag{4.2}$$

Sendo I uma imagem e B um elemento estruturante.

A Figura seguinte ilustra o processo de extração de etiquetas e falhas na digitalização. A imagem original apresentada na Figura 16-a é submetida a um limiar, ilustrado na Figura 16-b. Na próxima etapa, o limiar obtido passa pelo processo de abertura e fechamento, ilustrado em 16-c. Após esta etapa, o *array* da imagem original é multiplicado com a máscara, gerando a imagem final, ilustrada em 16-d.

4.2.2. Remoção do Músculo Peitoral

O músculo peitoral deve ser removido, pois não possui nenhuma informação útil ao processo de segmentação. Nesta etapa, foi usada a técnica de crescimento de regiões, que consiste em agregar *pixels* com propriedades similares em regiões. A técnica se inicia com a implantação de um *pixel* semente na região desejada. A partir daí, este *pixel* semente seleciona, através de um limiar atribuído e realiza o crescimento da vizinhança, agregando os *pixels* próximos que possuam atributos similares aos da semente. O processo continua até que se atinja uma condição de parada pré-estabelecida pelo usuário, como por

exemplo, um determinado nível de cinza ou uma distância específica (GONZALEZ, WOODS, 2007).

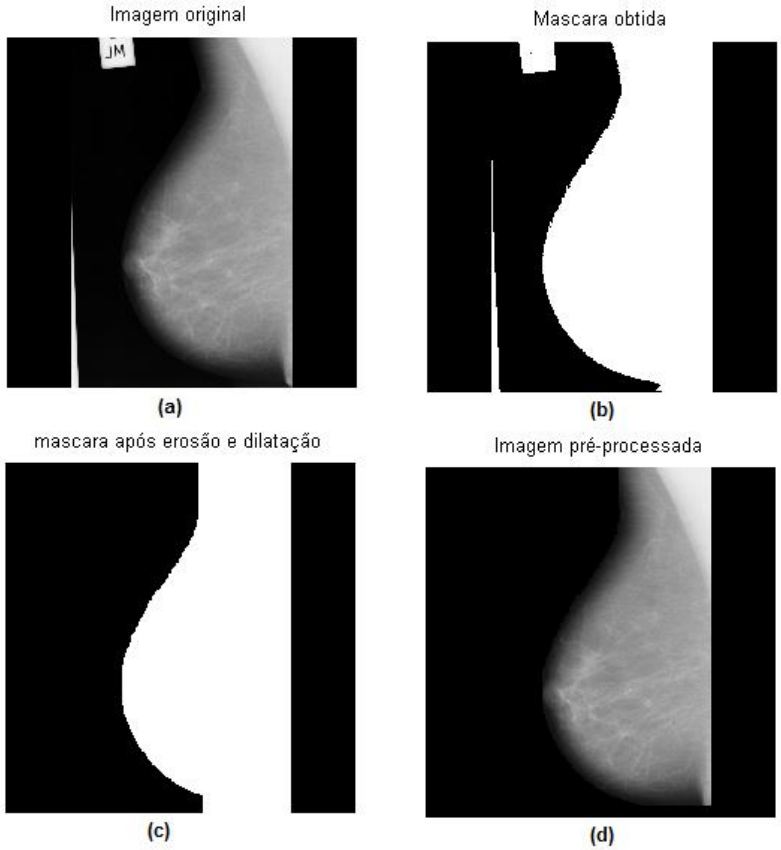


Figura 16 – Etapas do pré-processamento para remoção de etiquetas e falhas na digitalização, utilizando *thresholding* e operadores morfológicos

A Figura 17 ilustra a remoção do músculo peitoral da Figura 16 através da técnica de crescimento de regiões.

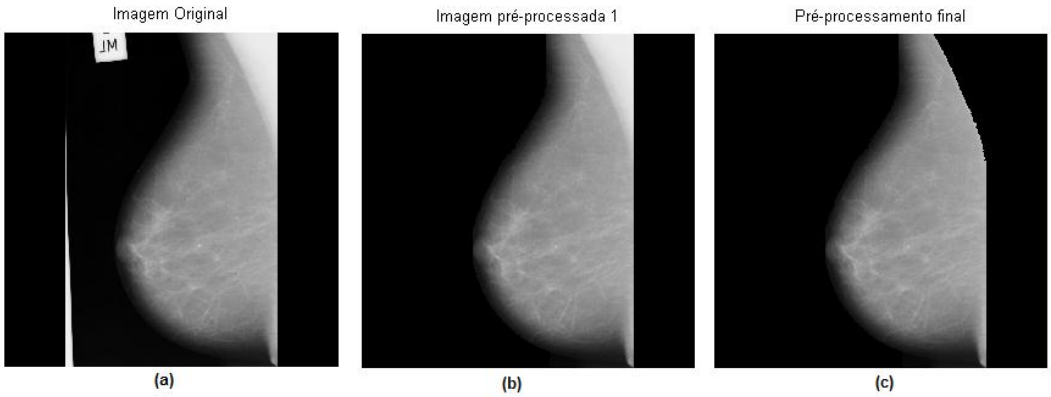


Figura 17 – A imagem original (17-a) sofreu abertura e fechamento, resultando na imagem pré-processada 1 (17-b). Em seguida foi realizada a segmentação por crescimento de regiões, para remoção do músculo peitoral, ilustrado na imagem Pré-processamento final (17-c).

4.3. Segmentação

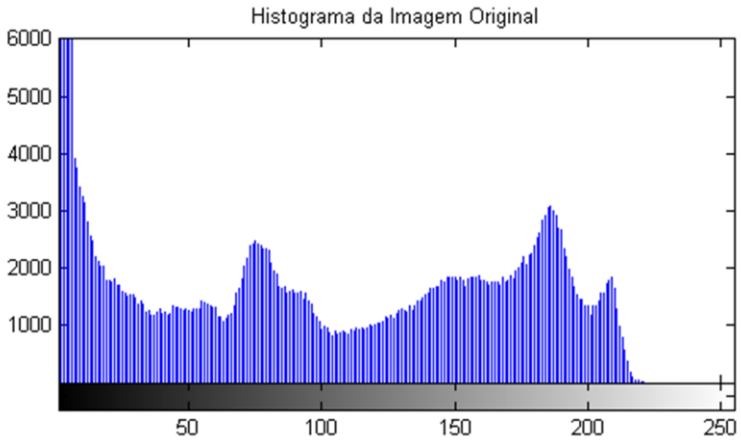
A segmentação tem o objetivo de identificar e isolar somente as regiões com maiores possibilidades de conter alguma espécie de lesão, que neste trabalhão são chamadas de Regiões Suspeitas, do inglês *Region of Suspicious* (ROS).

Foram utilizadas duas técnicas para a segmentação das ROS. Equalização Adaptativa de Histograma com Limitação de Contraste (CLAHE) e Alargamento de Contraste.

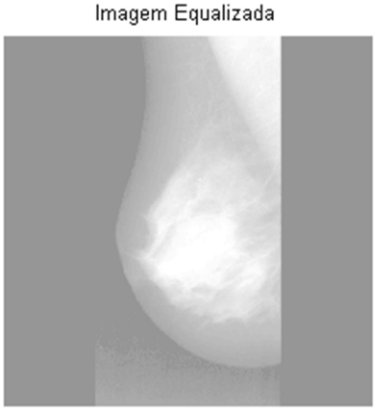
4.3.1. Equalização Adaptativa de Histograma com Limitação de Contraste (CLAHE)

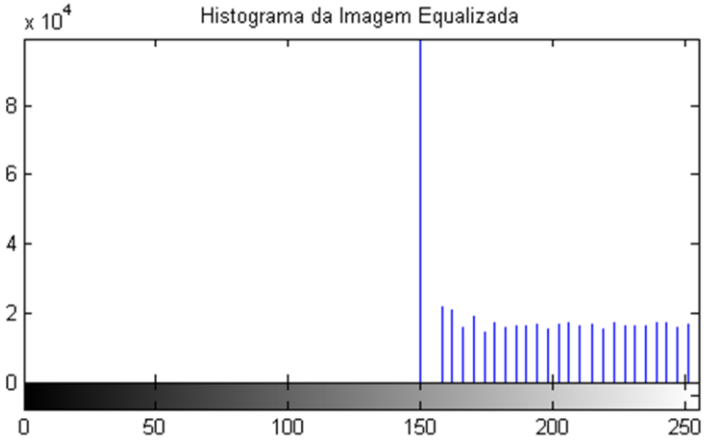
Conforme visto na seção 2.6.1, a CLAHE divide as imagens em pequenas regiões e faz a equalização individual de cada região, bem diferente da equalização tradicional, que utiliza a região inteira. Como resultado, tem-se uma imagem equalizada com o contraste local ampliado e limitado, para evitar amplificação de ruído e saturações. Assim, a imagem resultante apresentará mais detalhes.

A Figura 18 ilustra em resumo as aplicações descritas neste tópico. A Figura 18-a ilustra uma mamografia com tecido mamário extremamente denso, de difícil interpretação e seu respectivo histograma, em 18-b. A mamografia é então equalizada pelo processo tradicional e é mostrado seu histograma (18-c e 18-d). Finalmente, a mamografia é equalizada através de CLAHE (18-e), onde é claramente visível a variação do contraste e a riqueza de detalhes, comparada com 18-a e 18-b. Seu histograma é mostrado em 18-f.

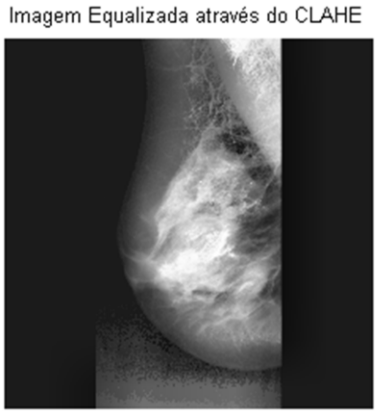


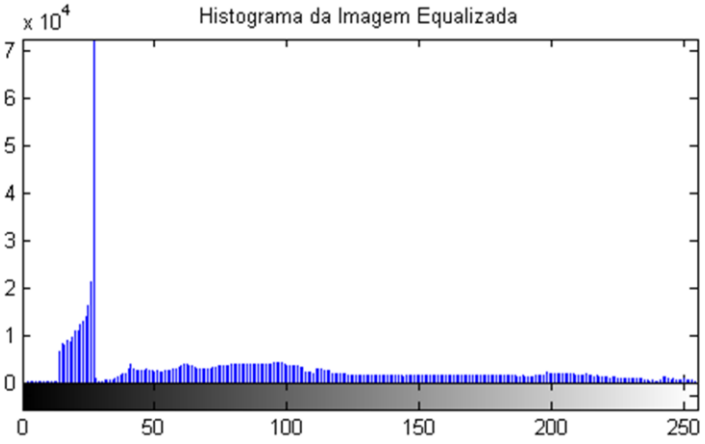
18.a **18.b**





18.c **18.d**





18.e **18.f**

Figura 18– Ilustração da comparação entre a equalização de histograma clássica, e a equalização adaptativa com limitação de contraste (CLAHE)

4.3.2. Alargamento de Contraste

O Alargamento de Contraste, conforme descrito na seção 2.6.2, expande uma determinada faixa de níveis de intensidade, sendo que a expansão (alargamento) ocorre somente no intervalo determinado pelos parâmetros da função de transformação linear por partes.

Neste trabalho, o alargamento foi utilizado para amplificar as regiões de maior contraste e reduzir a intensidade nas regiões de menor contraste. Dessa forma, apenas as regiões suspeitas de lesão foram isoladas. A Figura 19 ilustra todas as etapas desenvolvidas até o presente momento. A Figura 19-a ilustra uma mamografia extremamente densa, de difícil identificação de tecido lesionado. A Figura 19-b ilustra a etapa de pré-processamento observada na seção 4.2. Após esta etapa, foi realizada a etapa de equalização adaptativa com limitação de contraste (CLAHE), ilustrada em 19-c. Após a aplicação da Equalização Adaptativa de Histograma com Contraste Limitado, foi aplicado na imagem resultante alargamento de contraste (Figura 19-d), utilizando uma função de transformação similar à mostrada na Figura 20. Observa-se que apenas a região de maior intensidade da Figura 19-a é alargada, pois somente tal região possui contraste que pôde ser mapeado segundo a função de transformação aplicada. Dessa forma, é possível extrair somente a região de interesse para a segmentação, chamada de Região Suspeita, que será discriminada entre normal e anormal, através da etapa de classificação. Para tanto, é necessário extrair e selecionar características relevantes, que possam diferenciar a que classe a ROS encontrada pertence. As próximas etapas da metodologia buscam extrair e selecionar as características mais relevantes, e classificar as ROS perante as suas classes.

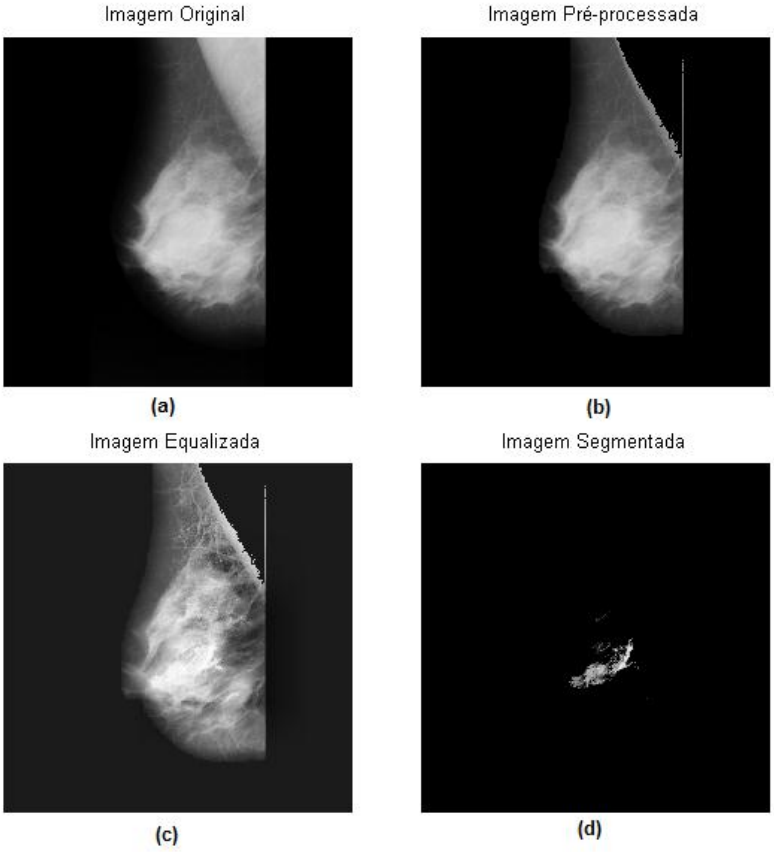


Figura 19 – Imagem Original (19-a). Pré-Processamento (19-b). Equalizado através de CLAHE (19-c). Alargamento de Contraste, apresentando a segmentação final (19-d).

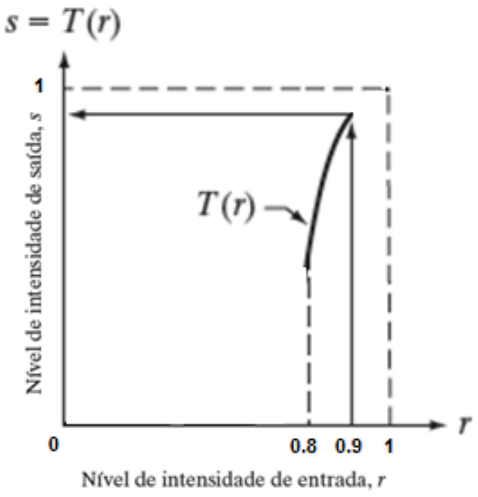


Figura 20 – Função de transformação aplicada no exemplo da Figura 19-d, sendo os valores de intensidade normalizados entre "0" e "1".

4.3.3. Extração de Características

O objetivo desta etapa é obter características a partir das Regiões Suspeitas (ROS) que o representem de forma mais descritiva e discriminativa entre os diferentes subgrupos de dados. As características selecionadas devem garantir que as ROS sejam corretamente classificadas como câncer e não câncer.

Nesta etapa foi utilizada a Análise de Componentes Independentes (ICA), vista na seção 2.6.3, onde a matriz **X** da equação (2.4) é representada usando as ROS encontradas na etapa da segmentação, re-escalonadas para um tamanho de *30 x 30pixels* e transformadas em um vetor dimensional, de dimensão:

$$P = 1\ x\ 900 \tag{4.3}$$

Cada amostra representa uma linha da matriz de mistura. A matriz **X** é representada por amostras na dimensão de $P = 1\ x\ 900$. Então, cada linha da matriz **A** corresponde a uma ROS e cada coluna corresponde a um peso atribuído para a imagem base, ou seja, um parâmetro de entrada para o classificador (CAMPOS et al, 2007; CHRISTOYIANNI, 2002).

Usando o algoritmo *FastICA* (MARCHINI, 2004)e a matriz **X**, foi obtida a matriz de funções bases **A**, que contém as características de cada amostra.

4.3.4. Seleção das Características Mais Significantes

O objetivo desta etapa é selecionar as características que melhor represente os dados gerados a partir da etapa de extração de características. Caso todos os dados gerados sirvam de entrada para um classificador, o resultado pode ser insatisfatório, com baixa acurácia e grande esforço computacional. Nesta etapa, foi utilizada a técnica de seleção de características por Máxima Relevância e Mínima Redundância (mRMR), visto na seção 2.6.4.1.

4.3.5. Classificação

A etapa de classificação vai analisar o vetor de características de cada Região Suspeita, reduzido através da técnica de mRMR e vai atribuir uma classificação, ou seja, vai rotular como normal ou anormal.

Para aumentar a confiabilidade do resultado do classificador, foi utilizada a técnica estatística de validação cruzada *10-fold-cross validation* (KIRALJ, FERREIRA, 2009), onde o conjunto de dados é dividido igualmente em 10 subconjuntos, o treino efetua-se concatenando 9 subconjuntos e a classificação usando o subconjunto restante. As fases de treino e teste são depois repetidas 10 vezes, permutando-se circularmente os subconjuntos. A acurácia final é calculada usando a média das acurácias de cada fase.

4.3.6. Avaliação do Método de Classificação

Sensibilidade, Especificidade e Acurácia são as medidas mais utilizadas para descrever um sistema de diagnóstico. Sensibilidade (S) é a proporção de verdadeiros positivos que são corretamente identificados pelo teste e é definida por S = VP/(VP+FN). Especificidade (E) é a proporção de verdadeiros negativos que são corretamente identificados no teste e é dada por E= VN/(VN+FP). Onde FN é Falso Negativo, FP Falso Positivo, VN Verdadeiro Negativo e VP Verdadeiro Positivo. Acurácia é a proporção de acertos, ou seja, o total de verdadeiros positivos e verdadeiros negativos em relação à amostra estudada (PEREIRA, 2008).

Também é possível avaliar um teste por meio de gráficos e mesmo usar esse procedimento para decidir qual o ponto de corte de uma distribuição, para definir seus valores de sensibilidade e especificidade. A curva ROC, do inglês *Receiver Operating Characteristic*, é uma forma de combinar valores de sensibilidade e especificidade, ou de

valores falso-positivos e falso-negativos e visualizá-los através de um gráfico. A Figura 21 ilustra uma curva ROC. Quanto mais a curva se afasta do canto superior esquerdo, mais ineficiente é o diagnóstico, e quanto mais próximo, mais eficiente (PEREIRA, 2008).

Outra medida importante que deve ser analisada na curva ROC é a Área sob a Curva, do inglês *Area under Curve*–AuC, que é uma medida que também serve como análise de desempenho.

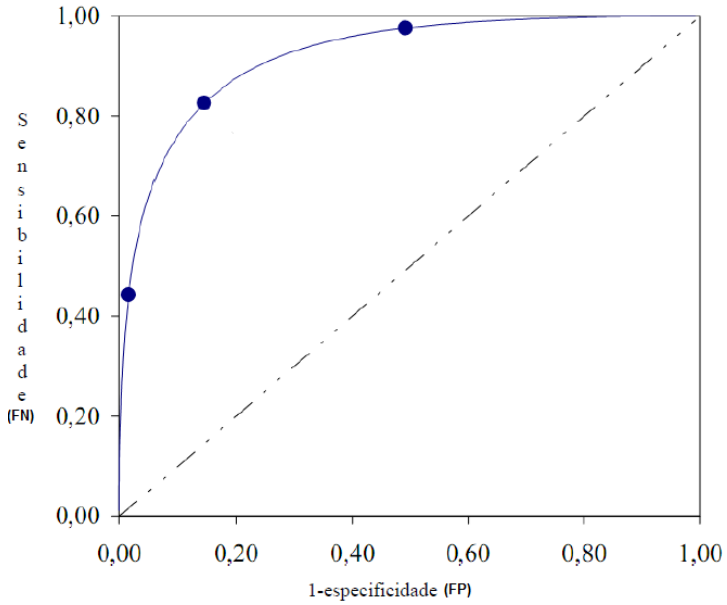


Figura 21 – Curva ROC

5. Resultados e Discussões

Uma série de testes foi realizada para avaliar o método proposto. Este capítulo apresenta e discute os resultados obtidos nas abordagens utilizadas.

5.1. Utilização da Base de Dados MIAS

Conforme exposto na seção 4.1, a base de dados utilizada neste trabalho possui 322 mamogramas, sendo que existem 122 casos anormais, seja benigno ou maligno. É importante lembrar que em algumas mamografias foram encontrados mais de um

caso. Destes 122 casos, 38 foram encontrados em mamas que possuem tecido gorduroso-glandular e em 40 que possuem tecido denso. Neste trabalho, foram utilizadas mamografias que possuem lesões de difícil rastreio, ou seja, aquelas encontradas em mamografias que o parênquima denso dificulta a detecção. Sendo assim, foram selecionados todos os casos com tecido gorduroso-glandular (38 casos) e denso (40 casos), totalizando um conjunto de 78 casos. A base de dados também informa a localização *xy* da lesão em cada mamografia com algum tipo de achado, seja ele benigno ou maligno (exceto calcificações). Entretanto, nas mamografias Mdb212 e Mdb214, a localização *xy* da lesão está equivocada, fora da mamografia. Sendo assim, tais mamografias foram excluídas, e o conjunto final para o teste do método foi realizado com 76 casos.

5.2. Segmentação Utilizando CLAHE e Alargamento de Contraste

Logo após a etapa de pré-processamento, que retirou placas de identificação, imperfeições na digitalização e o músculo peitoral de cada mamografia, foi iniciada a etapa de segmentação.

Primeiramente, cada mamografia foi submetida a uma Equalização Adaptativa de Histograma, com Limitação de Contraste (CLAHE), vista na seção 2.6.1. Nesta etapa, a mamografia foi dividida em regiões contextuais, de tamanho 128 x 256 *pixels* e o valor do *cliplimit* foi ajustado para $\alpha=0,2$.

Após esta etapa, cada mamografia passou por uma transformação de alargamento de contraste, com função de transformação similar a da Figura 21, onde somente os valores de intensidade compreendidos entre 0.8 e 0.9 foram alargados.

A partir daí, a etapa de segmentação selecionou um total de 500 regiões suspeitas (ROS), sendo 74 anormais (maligno ou benigno) e 426 normais. Dos 76 casos anormais (maligno ou benigno), o procedimento de segmentação não conseguiu rastrear 2, e encontrou os 74 restantes. Através do resultado encontrado, pode-se afirmar que a taxa de acerto, na etapa de segmentação, foi de 97.36%.

Seja taxa de Falsos Positivos por Imagem (FPI) calculada por:

$$FPI = \frac{\sum_{i=1}^{n} i_{FP}}{N} \tag{5.1}$$

Sendo N o numero de casos, e FP, os falsos positivos.

Através da equação 5.1, extraiu-se a FPI=5,6.

Todas as 500 ROS encontradas pelo procedimento de segmentação foram utilizadas como dados de entrada para um algoritmo de extração de características distintas para as ROS normais e anormais.

5.3. Extração de Características

A etapa de extração de características foi efetuada utilizando a técnica de Análise de Componentes Independentes (ICA), vista na seção 2.6.3.

Conforme a seção 4.3.3, cada ROS foi re-escalonada para um tamanho de 30x30 *pixels* e transformadas em um vetor linha de dimensão 1 x 900. A matriz **X**, do modelo de ICA é composta pelas ROS, de dimensão 1x900, sendo cada linha uma amostra, totalizando uma matriz com dimensão 500 x 900.

Usando o algoritmo *FastICA* (MARCHINI, 2004) e a matriz **X**, foi obtida a matriz de funções bases **A**, que contém as características de cada amostra.

5.4. Seleção das Características Mais Significantes

Com o objetivo de reduzir o vetor de característica de cada amostra, foi utilizado o algoritmo de Máxima Relevância e Mínima Redundância (mRMR), visto em 2.6.4.1.

Foram realizados testes para gerar vetores com 30, 20 e 10 características, sendo que cada vetor gerado foi testado com o classificador de Máquinas de Vetor de Suporte (SVM), a fim de encontrar o vetor de características com melhor desempenho.

5.5. Classificação

As 500 amostras foram divididas em 10 subconjuntos, de 50 amostras cada, com o objetivo de realizar o teste de validação cruzada *10-fold cross validation*. O classificador de Máquinas de Vetor de Suporte (seção 2.6.5) realizou o treino concatenando 9 conjuntos de 50 amostras, e testando com o conjunto restante. A acurácia final foi calculada através da média da acurácia de cada fase.

A Tabela 3 mostra a média dos indicadores de desempenho obtidos através do método *10-fold cross validation*.

Tabela 3 – Desempenho do classificador para cada vetor de característica

Características	VP	FP	FN	VN	(%)			AUC
					Especificidade	Sensibilidade	Acurácia	
10	**60**	**0**	**14**	**426**	**100**	**81.88**	**97.20**	**0.905**
20	**50**	**0**	**24**	**426**	**100**	**67.56**	**95.20**	**0.83**
30	**41**	**0**	**33**	**426**	**100**	**55.4**	**93.4**	**0.77**

Observou-se que o melhor resultado encontrado foi utilizando vetores com 10 características, obtendo acurácia de 97.2 %, sensibilidade de 81.88% e especificidade de 100%. A tabela 4 ilustra o resultado em cada iteração do teste *10-fold cross validation*.

Tabela 4 – Desempenho do teste *10-fold cross validation* para o melhor resultado obtido

Iterações	Especificidade	Sensibilidade	Acurácia
1	100	75	94
2	100	86.66	99
3	100	86.36	98
4	100	86.20	99
5	100	80.55	95
6	100	81.39	98
7	100	78	95
8	100	81.03	98
9	100	81.81	98
10	100	81.08	97
Média	**100**	**81.88**	**97.20**

A Figura 22 ilustra a curva ROC do melhor resultado obtido.

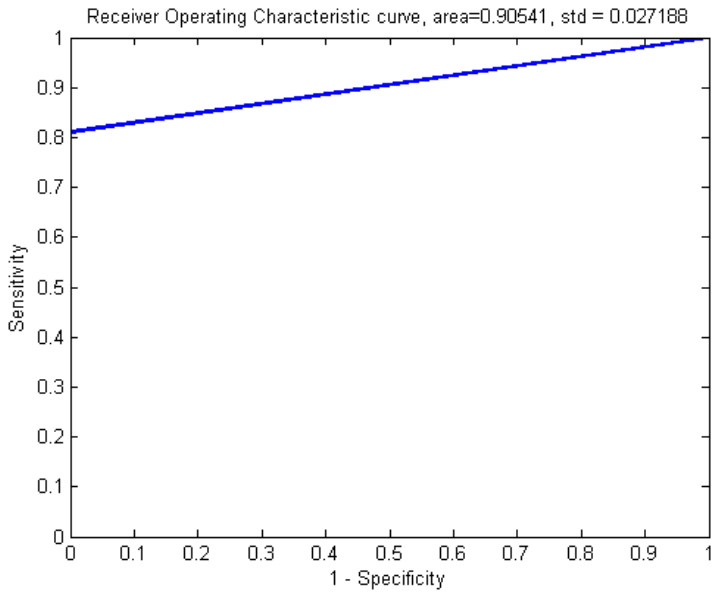


Figura 22 – Curva ROC do resultado obtido com 10 características

A etapa de segmentação consegue apenas encontrar regiões suspeitas de lesão, e devido ao tecido denso ser de difícil segmentação, a taxa de falso-positivos obtida foi elevada. Entretanto, a qualidade da extração e da seleção de características (seções 2.6.3 e 2.6.4) aplicadas no trabalho separou satisfatoriamente todas as ROS falso positivos das

ROS verdadeiro positivos.

Dos 76 casos anormais (benigno e maligno) submetidos à etapa de segmentação, 2 foram perdidos. Um dos fatores que podem estar associados a tal perda, é o erro de posicionamento das mamas, no momento do exame, que pode escurecer uma lesão. Dessa forma, com a lesão escura, e outras estruturas fibroglandulares da mama mais claros, não é possível obter melhoramento da região onde a lesão se encontra. Neste caso, evidencia-se a indispensável opinião do radiologista, que pode visualizar uma lesão que o sistema CAD não encontrou.

O melhor resultado encontrado, com um vetor de 10 características, pode ser justificado, devido ao menor esforço que o classificador SVM teve para discriminar as classes normais e anormais. Com um vetor menor o custo computacional diminuiu e acurácia aumentou, comparado ao pior resultado, onde o vetor de características tinha dimensão 1 x 30.

A curva ROC apresentou AuC=0.90, que demonstra que o classificador atingiu um bom desempenho. Também se pode levar em consideração, que devido às técnicas de extração (2.6.3) e seleção de características (2.6.4), o classificador SVM conseguiu separar as amostras de cada classe, com uma margem de erro de apenas 2.8%.

6. Conclusão

Nesta Tese, foi desenvolvido um método de detecção de Câncer, baseado em Equalização Adaptativa de Histograma, com Limitação de Contraste e Alargamento, para segmentar as regiões suspeitas de conterem algum tipo de lesão, e Análise de Componentes Independentes para extração de características e Máquinas de Vetor de Suporte (SVM), para a classificação final.

A etapa de segmentação obteve uma taxa de acerto de 97.36%, encontrando 74 lesões, das 76 analisadas. Em seguida a etapa de classificação analisou 500 ROS encontradas na segmentação, e obteve melhor resultado com um vetor de 10 características, atingindo uma acurácia de 97.2%, com sensibilidade de 81.88% e especificidade de 100%.

A taxa de falso positivo por imagem (FPI) pode ser considerada elevada, se comparada com a taxa de diversos algoritmos de segmentação já existentes. Entretanto, no caso apresentado neste trabalho, a segmentação foi realizada somente no tipo de mamografia mais difícil de segmentar, justificando a FPI elevada, na ordem de 5.6.

O software SADIM (Sistema de Auxilio de Diagnóstico em Imagens Mamográficas) desenvolvido pelo autor e sua equipe, só foi testado com mamas gordurosas, de complexidade baixa de segmentação e classificação. A metodologia proposta neste trabalho vai ser adicionada no software SADIM, para aumentar sua eficácia e viabilizar seus testes em hospitais e clínicas de radiologia.

Por fim, o presente trabalho abre a possibilidade para utilização em outras bases de dados, ou em análise de outros tipos de lesões, tais como nódulo pulmonar, nódulos encefálicos, etc.

REFERÊNCIAS

ACR: Breast Imaging Reporting and Data Systems (BI-RADS) 5nd edition. Reston: American College of Radiology, 2003.

ARONS, B. A review of cocktail party, Cambridge, MA: MIT laboratory, 1990

BOYD, N.F.; BYNG, J.W.; JONG, R.A.; FISHELL, E.K.; LITTLE, L.E.; MILLER, A.B.; LOCKWOOD, G.A.; TRITCHLER, D.L.; YAFFE, M.J. Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian national breast screening study. Journal of the National Cancer Institute, v. 87, p.670-675, 1995.

BOYD NF, ROMMENS JM, VOGT K, et al. Mammographic breast density as an intermediate phenotype for breast cancer. Lancet Oncol 2005;6:798-808.

BOYD NF, BYNG JW, JONG RA, et al. Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study. J Natl Cancer Inst 1995; 87:670-5

BOYD FN, GUO H, MARTIN LJ, et al. Mammographic Density and the Risk and Detection of Breast Cancer. The New England Journal of Medicine. 353;3. January 18, 2007.

BRAY F, McCARRON P, Parkin DM. The changing global patterns of female breast cancer incidence and mortality. Breast Cancer Res. 2004;6(6):4. 229-39.

BREASTCANCER. Types of Breast Cancer. Breastcancer.org. Disponível em: http://www.breastcancer.org/symptoms/types. Data de acesso: 03/04/2013

BREM RF, HOFFMEISTER JW, RAPELYEA JA, et al. Impact of breast density on computer-aided detection for breast cancer. AJR Am J Roentgenol. 2005;184:439–44.

BISHOP, C. M. Pattern Recognition and Machine Learning. New York: Springer, 2006.

BIRD, R. WALLACE, T. and YANKASKAS, B. "Analysis of cancer missed at screening mammography," *Radiology*, vol. 184, pp. 613–617, 1992.

BYRNE C, SCHAIRER C, BRINTON LA, et al.Effects of mammographic density and benign breast disease on breast cancer risk (United States). Cancer Causes Control 2001;12:103-10.

BYRNE C, SCHAIRER C, WOLFE J, et al. Mammographic features and breast cancer risk: effects with time, age, and menopause status. J Natl Cancer Inst 1995;87: 1622-9.

CALAS MJG, GUTFILEN B, PEREIRA WCA. CAD e mamografia: por que usar esta ferramenta? Radiol Bras. 2012 Jan/Fev;45(1):46–52.

CANCER COUNCIL. Ductal Carcinoma *in situ*.Câncer Council Victoria. Disponível em: http://www.cancervic.org.au/preventing-cancer/attend-screening/breasts-health/ductal-carcinoma-in-situ. Data de acesso: 03/04/2013

CAMPOS LFA, LEMOS ECM, SILVA LCO, COSTA D D., BARROS A K. Segmentation and Classification of Breast Cancer Using Independent Component Analysis, Texture Features and Neural Networks. Workshop de Informática Médica. 2011. Disponível em http://143.107.58.177/cecas/sites/default/files/wim%202011/WIM_Sessao_2_Artigo_2_Campos.pdf

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

CHOI, S. et al. Blind Source Separation and Independent Component Analysis: A Review. Neural Information Processing - Letters and Reviews, 6, n. 1, January 2005.

CHRISTOYIANNI I., KOUTRAS A., KOKKINAKIS G., "Computer aided diagnosis of breast cancer in digitized mammograms", Comp. Med. Imag. and Graph., 26:309-319, 2002.

COSTA, D. D.; CAMPOS, L. F. A. ; BARROS, A. K. Classification of breast tissue in mammograms using efficient coding. In *BioMedical Engineering OnLine* 2011. Avaliable at. http://www.biomedical-engineering-online.com/content/10/1/55

DING C., PENG H., "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," Proc. Second IEEEComputational Systems Bioinformatics Conf., pp. 523-528, Aug. 2003.

GILS CH, OTTEN JD, VERBEEK AL, HENDRIKS JH. Mammographic breast density and risk of breast cancer: masking bias or causality? Eur J Epidemiol 1998; 14:315-20.

GIERACH LG, ICHIKAWA L, KERLIKOWSKE K, et al. Relashionship Between Mammographic Density and Breast Cancer Death in the Breast Cancer Surveillance Consortium. Journal National of Cancer Institute. June 2012

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

HARVEY JA, BOVBJERG VE. Quantitative assessment of mammographic breast density: relationship with breast cancer risk.Radiology 2004; 230:29-41.

HAYKIN, S. Neural Networks: A comprehensive foundation. 2. ed. [S.l.]: Prentice Hall, 1999.

HO WT, LAM PW. Clinical performance of computer- assisted detection (CAD) system in detecting carcinoma in breasts of different densities. Clin Radiol. 2003;58:133–6.

HYVÄRINEN, A.; OJA, E. Independent Component Analysis: Algorithms and Application. Neural Networks, 13, n. 4-5, 2000.

HYVÄRINEN, A.; KARHUNEN, J.; OJA, E. Independent Component Analysis. New York: Wiley, 2001.

INCA. Instituto Nacional do Câncer. Programa Nacional de Controle do Câncer de Mama. Disponível em: http://www2.INCA.gov.br/wps/wcm/connect/acoes_programas/site/home/nobrasil/programa_controle_cancer_mama/ Data de acesso: 02/04/2013

INCA. Instituto Nacional do Câncer. Estimativa 2012/Incidência de Câncer no Brasil. Ministério da Saúde. Instituto Nacional do Câncer. Disponível em: www.INCA.gov.br. Data de acesso: 02/04/2013

INCA. Instituto Nacional do Câncer. ABC do câncer : abordagens básicas para o controle do câncer / Instituto Nacional de Câncer José Alencar Gomes da Silva, Coordenação Geral de Ações Estratégicas, Coordenação de Educação ; organização Luiz Claudio Santos Thuler. – 2. ed. rev. e atual.– Rio de Janeiro : INCA, 2012. 129 p.

KARSSEMEIJER N, OTTEN JDM, VERBEEK ALM, et al. Computer-aided detection versus independent double reading of masses on mammograms. Radiology.2003;227:192–200.

KERLIKOWSKE K, GRADY D, BARCLAY J, SICKLES EA, ERNSTER V. Effect of age, breast density, and family history on the sensitivity of first screening mammography.JAMA. 1996; 276:33-8.

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

MCCORMACK VA, SANTOS SILVA I. Breast density and parenchymal patterns as markers of a breast cancer risk: a meta-analysis. Cancer Epidemiol 2. Biomarkers Prev. 2006;15(6):1159-69.

MANDELSON TM, NINA O, PEGGY L. Breast Density as a Predictor of Mammography Detection: Comparison of Interval and Screen Detected Cancers. Journal of the National Cancer Institute, Vol 92, Nº 13, July, 2000.

MARCHINI J. L, HEATON C, RIPLEY B D. "FastICA algorithms to perform ICA and Projection Pursuit". (2004) Available at
http://www.stats.ox.ac.uk/~marchini/software.html

MARTINS L, JUNIOR GB, SILVA AC, PAIVA AC, GATTASS M. Detection of Masses in Digital Mammograms using K-means andSupport Vector Machine. Electronic Letters on Computer Vision and Image Analysis 8(2):39-50, 2009

MARQUES FILHO, Ogê; VIEIRA NETO, Hugo. *Processamento Digital de Imagens*, Rio de Janeiro: Brasport, 1999. ISBN 8574520098.

MAJID A, de PAREDES ES, DOHERTY RD, et al. Missed breast carcinoma: pitfalls and pearls. Radiographics. 2003;23:881–95.

MCCORMACK VA, dos SANTOS S I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta analysis. Cancer Epidemiol Biomarkers Prev 2006; 15: 1159-69.

MENOTTI D, SILVA WR. Classification of Mammograms by the Breast Composition. IPCV'12 - The 2012 International Conference on Image Processing, Computer Vision, and Pattern Recognition.

MOREIRA IC , AMARAL I , DOMINGUES I , CARDOSO A , CARDOSO MJ , CARDOSO JS .INbreast : Toward a Full-field Digital Mammographic Database. Academic Radiology.Volume 19, Issue 2, February 2012, Pages 236–248

OBENAUER S, SOHNS C, WERNER C, et al. Impact of breast density on computer-aided detection in full-field digital mammography. J Digit Imaging. 2006;19:258–63.

PAQUERAULT S, SAMUELSON FW, PETRICK N, et al. Investigation of reading mode and relative sensitivity as factors that influence reader performance when using computer-aided detection software. Acad Radiol. 2009; 16:1095–107.

PAPOULIS, A.; PILLAI, S. U.Probability, Random Variables and Sthocastic Processes. 4. ed. New York: McGraw-Hill, 2002.

PENG H, LONG F, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analisys and Machine Inteligence. Vol. 27, aug. 2005

PEREIRA, MG. Epidemiologia: Teoria e Prática. Rio de Janeiro: Guanabara, 2008

PERSSON I, THURFJELL E, HOLMBERG L. Effect of estrogen and estrogen-progestinreplacement regimens on mammographic breast parenchymal density. J Clin Oncol. 1997;15:3201-7.

PINKER K, PERRY N, MILNER S, MOKBEL K, DUFF S. Accuracy of breast cancer detection with full-fi eld digital mammography and integral computer-aided detection correlated with breast density as assessed by a new automated volumetric breast density measurement system. Breast Cancer Res. 2010; 12(Suppl 3): P4.

PIZER S.M, AMBURN E.O.P, J.D. Austin, et al, "Adaptive histogram equalization and its variations", CVGIP 39 (1987)355–368.

RAI RK, GOUR P, SINGH B. Underwater Image Segmentation using CLAHE Enhancement and Thresholding. International Journal of Emerging Technology and Advanced Engineering. Volume 2, Issue 1, January 2012.

SILVA WR, MENOTTI D. Classification of Mammograms by the Breast Composition. IPCV'12 - The 2012 International Conference on Image Processing, Computer Vision, and Pattern Recognition

SHEPHERD JA, KERLIKOWSKE K, MA L. Volume of Mammographic Density and Risk of Breast Cancer. Cancer Epidemology Biomarkers and Prevention: 20(7); 1473-82. 2011

SMOLA, A. J.; SCHÖLKOPF, B. Learning with Kernels. Cambridge: MIT Press, 2002.

SOHNS C, ANGIC B, SOSSALLA S, et al. Computer assisted diagnosis in full-field digital mammography– results in dependence of readers experiences. Breast J. 2010;16:490–7.

STONE J, DING J, WARREN RM, DUFFY SW. Predicting breast cancer risk using mammographic density measurements from both mammogram sides and views. Breast Cancer Res Treat 2010;124:551–4.

SUCKLING, J. et al. The mammographic images analysis society digital mammogram database. Experta Medica International Congress Series, 1069, 1994. 375-378.

SUGANTHI, G. VAIRA. SUTHA, J. *Classification of Breast Masses in Mammograms using Support Vector Machine*. In International Conference on Recent Advances and Future Trends in Information Technology (iRAFIT2012) Proceedings published in International Journal of Computer Applications® (IJCA)

SUNDARAMI M, RAMAR K, ARUMUGAMI N, PRABINI G. Histogram Based Contrast Enhacement for Mammogram Images. Proceedings of 2011 International

Conference on Signal Processing, Communication, Computing and Networking Technologies (ICSCCN 2011).

TABAR L, FAGERBERG G, CHEN HH, DUFFY SW, SMART CR, GAD A, et al. Efficacy of breast cancer screening by age. New results from the Swedish Two-County Trial. Cancer. 1995;75:2507-17.

TANG J, RANGAYYAN R M, XU J, NAQA I, YANG Y, Computer-Aided Detection and Diagnosis of Breast Cancer With Mammography: Recent Advances. IEEE Transactions on information technology in biomedicine, vol.13, n.02, march 2009.

TIEZZY, DG. Epidemiology of Breast Cancer. Revista Brasileira de Ginecologia e Obstetrícia. Editorial. 2009; 31(5):213(5)

URSIN G, MA H, WU AH, et al. Mammographic density and breast cancer in three ethnic groups. Cancer Epidemiol Biomarkers Prev 2003;12:332-8.

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

VERHEUS M, PEETERS PH, van NOORD PA, Grobbee DE, van Gils CH. No relationship between circulating levels of sex steroids 4 and mammographic breast density: the Prospect-EPIC cohort. Breast Cancer Res. 2007;9(4):R53.

VIGÁRIO, R. Extraction of ocular artifacts form ecg using independent components analysis, Eletroenceph. Clin. Neurophysiol., 103 (3) : 395-404, 1997.

WEXNER. Anatomy of the Breasts. The Ohio State University Wexner Medical Center. Disponível em: http://medicalcenter.osu.edu/patientcare/healthcare_services/breast_health/anatomy_of_the_breasts/Pages/index.aspx. Data de acesso: 02/04/2013

WOLFE JN, SAFTLAS AF, SALANE M. Mammographic parenchymal patterns and quantitative evaluation of mammographic densities: a case-control study. AJR Am J Roentgenol 1987; 148:1087-92.

WOLFE, J.N.: Risk for breast cancer development determined by mammographic parenchymal pattern. Cancer 37 (1976) 2486–2492

YANKASKAS BC, CLEVELAND RJ, SCHELL MJ, KOZAR R. Association of recall rates with sensitivity and positive predictive values of screening mammography. AJR Am J Roentgenol. 2001; 177:543-9.

YANG SK, MOON WK, CHO N, et al. Screening mammography-detected cancers: sensitivity of a computer-aided detection system applied to fullfield digital mammograms. Radiology. 2007; 244: 104–11.

YAGHJYAN L, COLDITZ GA, COLLINS L, et al. Mammographic Breast Density and Subsequent Risk of Breast Cancer in Postmenopausal Women According to Tumor Characteristics. Journal National of Cancer Institute. May 2011.