RENORBIO
Programa de Pós-Graduação em Biotecnologia

# Processamento e análise de sinais mamográficos na detecção do câncer de mama: diagnóstico auxiliado por computador (CAD)

DANIEL DUARTE COSTA

São Luís – MA
2012

REDE NORDESTE DE BIOTECNOLOGIA
Programa de Pós-Graduação em Biotecnologia
Ponto Focal – Universidade Federal do Maranhão

Daniel Duarte Costa

# Processamento e análise de sinais mamográficos na detecção do câncer de mama: diagnóstico auxiliado por computador (CAD)

Tese apresentada a Rede Nordeste de Biotecnologia, como requisito parcial para conclusão do curso de doutorado em Biotecnologia.

Área de Concentração: Biotecnologia em Saúde

Orientador: Prof. Dr. Allan Kardec Duailibe Barros Filho

São Luís – MA
2012

# Processamento e análise de sinais mamográficos na detecção do câncer de mama: diagnóstico auxiliado por computador (CAD)

Daniel Duarte Costa

Aprovada em _____ / _____ / _____

BANCA EXAMINADORA

_____

Prof. Dr. Allan Kardec Duailibe Barros Filho
Dr. em Engenharia Elétrica – RENORBIO/UFMA (Orientador)

_____

Prof. Drª. Maria do Desterro Soares Brandão Nascimento
Drª. em Medicina – RENORBIO/UFMA

_____

Prof. Drª. Maria Bethânia da Costa Chein
Drª. em Medicina - UFMA

_____

Profª. Drª. Alcione Miranda dos Santos
Drª. em Engenharia de Produção - UFMA

_____

Prof. Dr. Ewaldo Eder Carvalho Santana
Dr. em Engenharia Elétrica – UEMA

À Deus, fonte da vida.
À Ana Cláudia, minha esposa, pela compreensão e pela paciência.
À Manuela, minha querida filha, que antes mesmo de nascer já conseguiu dominar o meu coração.
Aos meus pais Lindemberg e Noêmia, pelo incentivo e carinho constantes.
Aos amigos, pelo apoio e companheirismo.
Aos professores e funcionários do Departamento de Engenharia Elétrica da Universidade Federal do Maranhão e da Rede Nordeste de Biotecnologia.

# AGRADECIMENTOS

À força criadora da vida.

Ao professor Dr. Allan Kardec Duailibe Barros Filho pelo apoio, paciência, competência e dedicação.

A todos os meus amigos do Laboratório de Processamento da Informação Biológica (PIB), principalmente ao Lúcio e ao Luís Cláudio que participaram diretamente no desenvolvimento deste projeto.

À RENORBIO e à UFMA por ter me dado à oportunidade de desenvolver este trabalho na modalidade de doutorado.

À minha esposa, Ana Cláudia, sempre compreensiva e amorosa.

Aos meus pais, Lindemberg e Noêmia, que sempre me apoiaram.

À minha irmã, Katiana, que me ajudou com a revisão deste texto.

À minha família ludovicense que sempre me acolheu muito bem, Tio Ademir, Ana Maria, Alice, Gabriel e Eduardo.

A um grande amigo, Geraldo Braz Junior, por estar sempre me ajudando.

A todos que, diretamente ou indiretamente, contribuíram para a elaboração deste trabalho.

"*Só atingiria sucesso na vida pela autodisciplina.*
*Apliquei-a até que meus desejos se realizassem.*"

Nicola Tesla

# RESUMO

COSTA, D. D. **Processamento e análise de sinais mamográficos na detecção do câncer de mama: diagnóstico auxiliado por computador (CAD).** 2012. 111p. Tese (Doutorado em Biotecnologia). Rede Nordeste de Biotecnologia - RENORBIO. São Luís.

O câncer de mama é a principal causa de morte por câncer na população feminina dos países ocidentais. Para melhorar a precisão do diagnóstico por radiologistas e fazê-lo de forma precoce, novos sistemas de visão computacional têm sido criados e melhorados com o decorrer do tempo. Alguns métodos de detecção e classificação da lesão em imagens radiológicas, por sistemas de diagnósticos por computador (CAD), foram desenvolvidos utilizando diferentes técnicas estatísticas. Neste trabalho, apresentam-se metodologias de sistemas CADs para detectar e classificar regiões de massa em imagens mamográficas, oriundas de duas bases de imagens: DDSM e MIAS. Os resultados mostram que é possível, através destas metodologias, obter uma taxa de detecção de até 96% das regiões de massa, utilizando a técnica de codificação eficiente com o algoritmo de agrupamento *k-means*, e classificar corretamente as regiões de massa em até 90% utilizando-se das técnicas de análise de componentes independentes (ICA) e análise discriminante linear (LDA). A partir destes resultados gerou-se uma aplicação *web*, denominada SADIM (Sistema de Auxílio a Diagnóstico de Imagem Mamográfica), que pode ser utilizado por qualquer profissional cadastrado.

Palavras-chave: processamento de imagens médicas; diagnóstico auxiliado por computador; mamografias – análise de imagens; codificação eficiente.

# ABSTRACT

COSTA, D. D. **Processamento e análise de sinais mamográficos na detecção do câncer de mama: diagnóstico auxiliado por computador (CAD).** 2012. 111p. Tese (Doutorado em Biotecnologia). Rede Nordeste de Biotecnologia - RENORBIO. São Luís.

Breast cancer is the leading cause of cancer death among women in Western countries. To improve the accuracy of diagnosis by radiologists and doing it so early, new computer vision systems have been developed and improved with the passage of time. Some methods of the detection and classification of lesions in mammography images for computer systems diagnostic (CAD) were developed using different statistical techniques. In this thesis, we present methodologies of CADs systems to detect and classify mass regions in mammographic images, from two image databases: DDSM and MIAS. The results show that it is possible by these methods to obtain a detection rate of up to 96% of mass regions, using efficient coding technique and K-means clustering algorithm. To classify regions in mass or non-mass correctly, was obtained a success rate up to 90% using the independent component analysis (ICA) and linear discriminant analysis (LDA). From these results generated a web application, called SADIM (Sistema de Auxílio a Diagnóstico de Imagem Mamográfica), which can be used by any registered professional.

Keywords: medical image processing, computer aided diagnosis, mammography - image analysis, efficient coding.

# LISTA DE ILUSTRAÇÕES

Figura 1: Ilustração de um mamógrafo, onde se pode ver a câmera, o feixe de raios-X e a placa para o filme onde fica armazenada a imagem mamográfica. Na mamografia, cada mama é comprimida horizontalmente e, em seguida, obliquamente, armazenando uma imagem de raios-X de cada posição. Adaptada de (BRANDÃO, 2012). ..... 27
Figura 2: (a) É a ilustração de uma mama com a sua representação mamográfica, onde se observa a apresentação destes tecidos na mamografia. Em (b) é ilustrado um exemplo de nódulo maligno, microcalcificações e nódulo benigno, respectivamente. Adaptada de (HARVARD MEDICAL SCHOOL - PORTUGAL PROGRAM, 2012). ..... 28
Figura 3: Mamografia com incidência craniocaudal. Dois tumores malignos devidamente marcados na mama esquerda de uma paciente de 65 anos. Fonte: banco de dados DDSM (HEATH et al., 1998). Identificação Volume: câncer 01 Caso: B-3027-1. ..... 30
Figura 4: Mamografia com a incidência médio-lateral oblíqua. Dois tumores malignos devidamente marcados na mama esquerda de uma paciente de 65 anos. Fonte: banco de dados DDSM (HEATH et al., 1998). Identificação Volume: câncer 01 Caso: B-3027-1 ..... 30
Figura 5: Contorno dos nódulos. (a) Contorno regular; (b) contorno lobulado; (c) contorno irregular; (d) contorno espiculado. ..... 31
Figura 6: Limite dos nódulos. (a) Limite definido; (b) limite parcialmente definido; (c) limite pouco definido. ..... 32
Figura 7: Densidade dos nódulos. (a) Nódulo denso; (b) nódulo isodenso. (c) nódulo com baixa densidade; (d) nódulo com densidade de gordura; (e) nódulo com densidade heterogênea. ..... 32
Figura 8: Sistema Visual Artificial. As principais etapas de um SVA podem ser dividas em cinco passos: aquisição, pré-processamento, segmentação, extração de características e reconhecimento e interpretação. Todas as etapas estão interligadas através de uma base de conhecimento. Adaptada de (FILHO; NETO, 1999). ..... 35
Figura 9: Equalização do histograma. Exemplo de uma equalização de histograma em uma imagem médica de escaneamento cerebral. Observa-se que houve um realce na imagem. Adaptada de (PRATT, 2001). ..... 36
Figura 10: Exemplo da aplicação de um filtro de média. (a) É a imagem original. Em (b) foi adicionado um ruído gaussiano à imagem original. Em (c) observa-se o resultado da filtragem utilizando uma máscara de média 5x5. Verifica-se que houve uma redução destes ruídos. Adaptada de (BOVIC; ACTON, 2001). ..... 36
Figura 11: Exemplo de aplicação de operadores morfológicos. Em (a) observa-se a imagem original. (b) É o resultado da operação de dilatação e (c) é o resultado da operação de erosão da imagem original. Constata-se que houve uma maior evidência em algumas áreas, dependendo do filtro aplicado. Adaptada de (PRATT, 2001). ..... 36
Figura 12: Modelo padrão para respostas das células simples de V1. O neurônio calcula uma soma ponderada da imagem no espaço e no tempo. O resultado é normalizado pelas respostas de unidades vizinhas e passada através de uma não linearidade ponto-a-ponto (CARANDINI; HEEGER; MOVSHON, 1997). Adaptada de (OLSHAUSEN; FIELD, 2004). ..... 40
Figura 13: Banco de filtros de wavelets de Gabor. Campos receptivos de uma célula complexa. Fonte: (LENNIE, 2003). ..... 41
Figura 14: Direção das componentes principais. (a) Representa o conjuto de dados originais. Após a PCA, em (b), é encontrado o vetor **Y1** que aponta na direção de maior variância dos dados e **Y2** que é

o vetor que aponta para a segunda maior variância, obedecendo o critério de ser ortogonal ao vetor **Y1**. Em (c) são ilustrados esses dados nestes novos eixos, descorrelacionando-os. .....42
Figura 15: Metodologia proposta em três passos para segmentação de massas em mamografias digitalizadas: aquisição de imagens; extração de características; agrupamento por *k-means*, nebuloso *c-means* e mapa auto-organizável. .....56
Figura 16: Passo a passo da remoção de ruídos e artefatos das mamografias. Em (a) tem-se a imagem mdb005 do banco de dados MIAS (SUCKLING et. al., 1994). Em (b) tem-se a imagem original filtrada pelo filtro de média. (c) É a imagem binarizada utilizando o limiar global do método de Otsu. (d) É a imagem binária após a operação de abertura. (e) Ilustra o passo da remoção de artefatos da mamografia. (f) Resultado deste processo de remoção de ruídos e artefatos. .....59
Figura 17: Oito filtros de Gabor utilizados para a filtragem das imagens. .....61
Figura 18: Oito filtros da PCA utilizados para a filtragem das imagens. .....61
Figura 19: Oito filtros da ICA utilizados para a filtragem das imagens. .....61
Figura 20: Convolução da imagem com os filtros gerados pela extração de características. Nesta figura a imagem original foi convoluída para cada um dos oito filtros da ICA, resultando assim em oito imagens que devem ser associadas pelos algoritmos de agrupamento. .....63
Figura 21: Localização das regiões de interesse. Em (a) tem-se a imagem original. (b) Ilustra o resultado do agrupamento realizado pelo *k-means*, isolando-se apenas o grupo da região do músculo. (c) É a região de interesse que foi obtida pela retirada da região do músculo peitoral maior. .....64
Figura 22: Imagem MDB005 do MIAS segmentada pelos três algoritmos de agrupamento e as três técnicas de codificação. .....68
Figura 23: Imagem MDB010 segmentada pelos três algoritmos de agrupamento e as três técnicas de codificação. .....69
Figura 24: Imagem MDB017 segmentada pelos três algoritmos de agrupamento e as três técnicas de codificação. .....70
Figura 25: Imagem MDB075 segmentada pelos três algoritmos de agrupamento e as três técnicas de codificação. .....71
Figura 26: Metodologia proposta em três passos para classificação de câncer de mama em imagens digitais. O primeiro passo é a aquisição de imagens, onde posteriormente serão divididas em dois grupos, treino e teste; a segunda etapa é a extração de características pelas técnicas de codificação: *wavelets* de Gabor, análise de componentes principais e análise de componentes independentes; o último passo é o de classificação pela análise discriminante linear, que dirá a que classe pertence o tecido analisado. .....77
Figura 27: Equalização do histograma de uma região de interesse. Observa-se que houve um realce na imagem após o procedimento de equalização do histograma. Fonte: Imagem editada a partir da imagem mdb005.pgm do banco de dados MIAS (SUCKLING et. al., 1994). .....78
Figura 28: *Wavelets* de Gabor escolhidas pelo algoritmo de busca. Dentre as 100 *wavelets* geradas, estas são as 50 mais significativas, conforme o algoritmo de busca desta metodologia. .....80
Figura 29: Componentes principais escolhidas pelo algoritmo de busca. Dentre as 1024 componentes geradas, estas são as 50 mais significativas, conforme o algoritmo de busca desta metodologia. .....81
Figura 30: 50 funções bases pelo modelo de codificação eficiente e selecionadas pelo algoritmo de busca. .....83
Figura 31: Média das acurácias obtidas pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em massa ou não-massa. .....86

Figura 32: Média das sensibilidades obtidas pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em massa ou não-massa. ..... 87
Figura 33: Média das especificidades obtidas pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em massa ou não-massa. ..... 87
Figura 34: Média dos valores preditivos positivos obtidos pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em massa ou não-massa. ..... 88
Figura 35: Média dos valores preditivos negativos obtidos pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em massa ou não-massa. ..... 89
Figura 36: Média das acurácias obtidas pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em benigno ou maligno. ..... 90
Figura 37: Média das sensibilidades obtidas pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em benigno ou maligno. ..... 90
Figura 38: Média das especificidades obtidas pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em benigno ou maligno. ..... 91
Figura 39: Média dos VPP obtidos pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em benigno ou maligno. ..... 92
Figura 40: Média dos VPN obtidos pelo método 10-fold cross validation e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em benigno ou maligno. ..... 92
Figura 41: Visão geral do sistema proposto por SILVA (2012). O usuário efetua o cadastro e envia a imagem através da internet para o servidor que irá realizar uma sugestão de diagnóstico e repassa a informação para um especialista emitir o diagnóstico final. Adaptada de (SILVA, 2012) ..... 96
Figura 42: Mapa de calor. Em (a) temos a imagem original, do banco de dados MIAS, da qual já foi retirado ruídos, artefatos e o músculo do peito. Em (b) temos a imagem original em forma de mapa de calor, quanto mais vermelho, maiores as chances daquela região conter algum tipo de nódulo. (c) São ilustradas apenas as regiões quentes que contenham pelo menos 66% de vermelho. Em (d) (e) e (f) são ilustradas respectivamente as regiões com 75%, 90% e 95% de vermelho. ..... 100

# LISTA DE TABELAS

Tabela 1: Classificação dos tumores conforme proposto pela União Internacional Contra o Câncer (UICC). Fonte: (MINISTÉRIO DA SAÚDE. SECRETARIA DE ATENÇÃO A SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2004a). ..... 21
Tabela 2: Sistemas de diagnóstico de câncer de mama auxiliado por computador. A primeira coluna é a referência do trabalho, a segunda é a metodologia utilizada, a terceira é o banco de imagens e a quarta é a quantidade de imagens utilizadas. As últimas colunas são os resultados apresentados, sendo que FLL é a fração de lesão localizada, FNL é a fração de não lesão localizada, ACC é a acurácia, SEN é a sensibilidade e ESP é a especificidade. ..... 23
Tabela 3: Variância de cada componente principal e a variância acumulada, na redução de dimensionalidade por PCA. ..... 60
Tabela 4: Quantidades de verdadeiros positivos (VP) e falsos positivos (FP) de um total de 30 imagens para cada técnica e seu respectivo algoritmo de agrupamento. ..... 66
Tabela 5: Avaliação da metodologia da segmentação através de suas medidas: fração de localização de lesão (FLL) e fração de não lesão localizada (FNL). ..... 66
Tabela 6: Quantidades de verdadeiros positivos (VP) e falsos positivos (FP) de um total de 35 imagens de mamas densas para cada técnica e seu respectivo algoritmo de agrupamento. ..... 72
Tabela 7: Avaliação da metodologia da segmentação através de suas medidas fração de localização de lesão (FLL) e fração de não lesão localizada (FNL) de um total de 35 imagens de mamas densas para cada técnica e seu respectivo algoritmo de agrupamento. ..... 72
Tabela 8: Síntese dos resultados obtidos. As duas primeiras linhas são referentes à metodologia aplicada no capítulo 4, onde o segundo resultado foi obtido exclusivamente por mamas densas como teste. Os dois últimos resultados são referentes à metodologia do capítulo 5, onde o primeiro resultado diferencia regiões de massa das regiões de não-massa e o último é a discriminação entre imagens benignas de malignas. ..... 96

# LISTA DE ABREVIATURAS E SIGLAS

ACC – Acurácia.

AEM – Autoexame das mamas.

BSS – *Blind source separation* (separação cega de fontes).

CAD – *Computer aided diagnosis* (diagnóstico auxiliado por computador).

CC – Crânio caudal.

DDSM - *Database for screening mammography*.

DNA – *Deoxyribonucleic acid* (ácido desoxirribonucleico).

ECG – Eletrocardiograma.

ECM – Exame clínico das mamas.

EEG – Eletroencefalograma.

ESP – Especificidade.

FCM – *Fuzzy c-means* (nebuloso *c-means*).

FLL – Fração de lesão localizada.

FNL – Fração de não lesão localizada.

GLCM – *Gray level co-ocurrence matrix* (matriz de co-ocorrência de tons de cinza).

ICA – *Independent component analysis* (análise de componentes independentes).

INCA – Instituto nacional do câncer.

LDA – *Linear discriminant analysis* (análise discriminante linear).

MAGIC-5 – *Medical Applications on a Grid Infrastructure Connection*-5 (Aplicações médicas em uma conexão de Infraestrutura de grade).

MCG – Magnetocardiograma.

MIAS – *Mammographic image analysis society*.

MLO – Médio lateral oblíquo.

MLP – *Multilayer Perceptron* (perceptron multicamada).

MMG – Mamografia.

MS – Ministério da Saúde.

PCA – *Principal component analysis* (análise de componentes principais).

ROI – *Region of interest* (região de interesse).

SEN – Sensibilidade.

SOM – *Self-organizing maps* (mapa auto-organizável).

SVA – Sistema visual artificial.

SVM – *Support vector machine* (máquina de vetor de suporte).

V1 – Área do córtex visual.

VPN – Valor preditivo negativo.

VPP – Valor preditivo positivo.

YGOPH – *Yaounde Gynaeco–Obstetric and Pediatric Hospital* (hospital ginecológico, obstétrico e pediátrico Yaounde).

# SUMÁRIO


**1. INTRODUÇÃO** ..... 17

**1.1. Diagnóstico Auxiliado por Computador** ..... 20

**1.2. Objetivo** ..... 25

**1.2.1. Objetivo Geral** ..... 25

**1.2.2. Objetivos Específicos** ..... 25

**1.3. Organização do trabalho** ..... 26

**2. MAMOGRAFIA E O CÂNCER DE MAMA** ..... 27

**3. FUNDAMENTOS TEÓRICOS** ..... 34

**3.1. Processamento digital de imagens** ..... 34

**3.2. Extração de características** ..... 38

**3.2.1. *Wavelets* Gabor** ..... 40

**3.2.2. Análise de Componentes Principais** ..... 42

**3.2.3. Análise de Componentes Independentes** ..... 44

**3.3. Algoritmos de Agrupamento** ..... 49

***3.3.1. K-means*** ..... 50

***3.3.2.* Nebuloso *c-means*** ..... 51

**3.3.3. Mapa auto-organizável** ..... 52

**3.4. Análise discriminante linear** ..... 54

**4. MÉTODO E RESULTADOS DA SEGMENTAÇÃO** ..... 55

**4.1. Introdução** ..... 55

**4.2. Aquisição de Imagens** ..... 57

**4.3. Extração de Características** ..... 60

**4.4. Agrupamento** ..... 62

**4.5. Avaliação do Método** ..... 64

**4.6. Resultados** ..... 65

**4.7. Discussões e Conclusões** ..... 73

**5. MÉTODO E RESULTADOS DA CLASSIFICAÇÃO** ..... 76

**5.1. Introdução** ..... 76

**5.2. Aquisição de Imagens** ..... 77

**5.3. Extração de Características** ..... 79

**5.4. Classificação** ..... 83

**5.5. Avaliação do Método** ..... 84

**5.6. Resultado** ..... 85
**5.7. Discussões e Conclusões** ..... 93
**6. CONSIDERAÇÕES FINAIS E PERSPECTIVAS FUTURAS** ..... 95
REFERÊNCIAS ..... 102
GLOSSÁRIO ..... 111

## 1. INTRODUÇÃO

O câncer de mama é o segundo tipo de câncer mais frequente no mundo, perdendo apenas para o câncer de pele do tipo não melanoma, e o mais comum entre as mulheres (MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2012). As pesquisas mostram que em 2012, 27,9% dos novos casos de câncer em mulheres serão de mama. No Brasil, são esperados aproximadamente 53.000 novos casos, com um risco estimado de 52,5 casos a cada 100 mil mulheres (MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2011). Na região Nordeste, o risco estimado é de 31,9 novos casos a cada 100 mil mulheres (MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2011).

O Instituto Nacional do Câncer (INCA) estima que no ano de 2012 a cada 9,5 minutos uma mulher será diagnosticada com câncer de mama no Brasil e a cada 42,8 minutos uma pessoa morrerá devido a esta enfermidade (MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2011).

O câncer de mama é resultado de alterações do DNA, que levam a uma proliferação celular desordenada (*AMERICAN CANCER SOCIETY*, 2011a). Quando há uma falha do mecanismo regulador que mantêm o equilíbrio entre o crescimento celular e o bem estar do organismo, ocorre a formação de massas, denominadas tumores ou neoplasias, as quais são classificadas como malignas ou benignas. As neoplasias benignas têm crescimento organizado, em geral lento, e o tumor apresenta contorno bem nítido. Na neoplasia maligna, o crescimento é rápido, desordenado e infiltrativo. Nos tumores malignos, suas células têm capacidade de se desenvolver em outras partes do corpo, fenômeno este denominado metástase (BAUER; IGOT; LE, 1980).

Além do fator gênero feminino, a idade é o elemento de risco mais importante no câncer de mama (AMERICAN CANCER SOCIETY, 2011a). A ameaça também aumenta com mutações genéticas hereditárias nos genes BRCA1, BRCA2 e p53, histórico pessoal ou familiar de câncer de mama, alta densidade no tecido mamário (uma medida mamográfica de quantidade de tecido glandular em relação ao tecido gorduroso de mama), biópsia confirmada de hiperplasia atípica e altas doses de radiação no peito como resultado de procedimentos médicos (THULER, 2003; *AMERICAN CANCER SOCIETY*, 2011b).

Os fatores de risco relacionados à vida reprodutiva da mulher, como a menarca precoce, nuliparidade (maior número de ovulações), primeira gestação tardia (acima dos 30 anos), anticoncepcionais orais, menopausa tardia e terapia de reposição hormonal, estão bem estabelecidos em relação ao desenvolvimento do câncer de mama (THULER, 2003; AMERICAN CANCER SOCIETY, 2011a, 2011b).

O melhor tratamento contra o câncer é a prevenção. As ações de prevenção primária objetivam diminuir a incidência na população, reduzindo o risco de surgimento de casos novos, ao prevenir a exposição aos fatores que levam ao seu desenvolvimento, interromper seus efeitos ou alterar as respostas do hospedeiro a essa exposição, impedindo que ocorra seu início biológico (THULER, 2003; MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2011). No entanto, esta prevenção ainda não é totalmente possível devido à variação dos fatores de risco e às características genéticas que estão envolvidas na etiologia.

A prevenção secundária tem por finalidade alterar o curso da doença, uma vez que seu início biológico já aconteceu, por meio de intervenções que permitam sua detecção precoce e seu tratamento oportuno. Para isso, deve haver clara evidência de que a doença em questão possa ser identificada em uma fase precoce, quando ainda não está clinicamente aparente, e que permita uma abordagem terapêutica eficaz, alterando seu curso ou minimizando os riscos associados com a terapêutica clínica (THULER, 2003). Quando o objetivo é detectar a doença precocemente estima-se que haja redução de até 90% na mortalidade por câncer, como o que ocorre nos programas de rastreamento para o câncer de mama (MINISTÉRIO DA SAÚDE. SECRETARIA DE ATENÇÃO A SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2007).

Os programas de rastreamento para câncer de mama têm como objetivo identificar mulheres que se encontram em estágio precoce da doença. Atualmente há três estratégias disponíveis para rastreamento do câncer de mama: autoexame das mamas (AEM), exame clínico das mamas (ECM) e mamografia (MMG) (MILLER, 2008; MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2009; *AMERICAN CANCER SOCIETY*, 2012).

O AEM é um exame que deve ser realizado mensalmente pela própria mulher, onde se deve procurar por protuberâncias, ondulações, checar a espessura da pele das mamas e

liberação de líquidos pelo mamilo. No entanto o INCA (MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2004, 2009) recomenda que este exame não substitua o ECM que deve ser realizado anualmente, principalmente em mulheres com mais de 40 anos, em uma consulta ginecológica. O médico deve examinar com as mãos a mama da paciente procurando encontrar sinais e sintomas de doenças. A MMG é um exame radiológico que examina o tecido da mama para detectar o câncer em fase inicial podendo descobrir alguns tipos de cânceres antes da paciente ou médico poder sentir o nódulo. Segundo o Ministério da Saúde (MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2004, 2009), este exame deve ser realizado no máximo a cada dois anos em mulheres com idade entre 50 a 69 anos. Estima-se que a cada 100 mulheres, com câncer de mama, o AEM detecta 26, o ECM detecta 45 e a MMG 71 (THULER, 2003).

De acordo com o INCA, a MMG tem sensibilidade entre 88% a 93,1% e especificidade entre 85% a 94,2%, e a utilização desse exame como método de rastreamento reduz a mortalidade em 25% (MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2007).

A sobrevida das pacientes é diretamente relacionada com o tamanho do tumor no diagnóstico inicial. O diagnóstico precoce não apenas influência o prognóstico, mas propicia cirurgia menos mutilante e com sobrevida comparável a intervenções cirúrgicas mais dramáticas e agressivas (BAUER; IGOT; LE, 1980).

A sobrevida média após cinco anos na população de países desenvolvidos tem apresentado um discreto aumento, cerca de 85%. Entretanto, nos países em desenvolvimento, a sobrevida fica em torno de 60%. (FERNANDES; NARCHI, 2007; MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2011). As taxas de mortalidade por câncer de mama continuam elevadas no Brasil, muito provavelmente porque a doença ainda seja diagnosticada em estágios avançados. Nesses estágios, a doença já evoluiu de forma que o organismo não é capaz de responder ao tratamento, que em geral é mutilante e causa maior sofrimento à mulher. Com base nos dados disponíveis de Registros Hospitalares, 50% dos tumores de mama, em média, são diagnosticados em estágios III e IV (COSTA; CASTIER, 2012). Na Tabela 1 estão caracterizadas as classificações dos estágios do câncer de mama com os seus respectivos tratamentos.

## 1.1. Diagnóstico Auxiliado por Computador

Sistemas de diagnóstico auxiliado por computador (CAD, do inglês *computer-aided diagnosis*) vêm sendo desenvolvidos por diversos grupos de pesquisa, visando auxiliar a detecção precoce do câncer de mama para que a paciente possa ter o tratamento adequado o mais cedo possível. É sabido que o melhor método de prevenção é o diagnóstico precoce, porque diminui a mortalidade e aumenta a eficácia do tratamento (MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2011). A sobrevida das pacientes é diretamente relacionada com o tamanho do tumor no diagnóstico inicial (BAUER; IGOT; LE, 1980). Portanto, um grande esforço tem sido feito para melhorar as técnicas de diagnóstico precoce através da MMG, que é um método simples, barato e acessível. Dessa forma, procura-se associar a heurística de identificação de padrões reconhecíveis baseada na percepção visual da imagem aos esquemas de CAD para extrair um diagnóstico.

O objetivo principal destes sistemas é fornecer aos médicos uma segunda opinião ou uma sugestão de diagnóstico. Os CADs contornam problemas que surgem da subjetividade de um laudo humano, tais como expectativas, pré-conceitos e cansaço que interferem no diagnóstico do radiologista. Buscam assim reduzir as taxas de falsos positivos, ou seja, casos em que a suspeita de malignidade leva a mulher à biópsia - exame altamente invasivo e de maior custo - para um resultado negativo. Vários estudos revelam que algoritmos automáticos de detecção são capazes de proporcionar um aumento de mais de 20% no total de acertos do radiologista e, com isso, consegue-se evitar biópsias desnecessárias (ZHANG; SANKAR; QIAN, 2002; BALLEYGUIER et al., 2005; SOHNS et al., 2010).

Para que isto ocorra, é importante desenvolver técnicas para detectar regiões suspeitas e reconhecer lesões. Vários métodos de diagnósticos de patologias em MMG têm sido desenvolvidos por diferentes grupos de pesquisadores, contribuindo para um diagnóstico mais confiável.

**Tabela 1: Classificação dos tumores conforme proposto pela União Internacional Contra o Câncer (UICC). Fonte: (MINISTÉRIO DA SAÚDE. SECRETARIA DE ATENÇÃO A SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2004a).**

| Estágios | Características | Tratamento |
|---|---|---|
| **Estágio 0** | É o chamado carcinoma *in situ* que não se infiltrou pelos ductos ou lóbulos, sendo um câncer não invasivo. O estágio zero significa que as células do câncer estão presentes ao longo da estrutura de um lóbulo ou um ducto, mas não se espalharam para o tecido gorduroso vizinho, isto é, as células cancerosas que ainda não invadiram os tecidos circundantes. | O tratamento padrão é uma lumpectomia com radioterapia ou uma mastectomia. |
| **Estágio I** | O tumor invasivo é pequeno (menos de 2 centímetros de diâmetro) e não se espalhou pelos linfonodos, isto é, o tumor que permaneceu no local no qual se originou sem disseminação para os linfonodos ou locais distantes. | O tratamento padrão é uma lumpectomia com radioterapia ou uma mastectomia com algum tipo de retirada de nódulo. A terapia hormonal, a quimioterapia e a terapia biológica também podem ser recomendadas após a cirurgia. |
| **Estágio IIa** | Qualquer das condições: o tumor tem menos que 2 centímetros e infiltrou linfonodos axilares; ou o tumor tem entre 2 a 5 centímetros, mas não atinge linfonodos axilares. | |
| **Estábio IIb** | Qualquer das condições: o tumor tem de 2 a 5 centímetros e atinge linfonodos axilares; ou o tumor é maior que 5 centímetros, mas não atinge linfonodos axilares. | |
| **Estágio IIIa** | Qualquer das condições: o tumor é menor que 5 centímetros e se espalhou pelos linfonodos axilares que estão aderidos uns aos outros ou a outras estruturas vizinhas; ou o tumor é maior que 5 centímetros, atinge linfonodos axilares os quais podem ou não estar aderidos uns aos outros ou a outras estruturas vizinhas. | O tratamento envolve cirurgia, provavelmente seguida de quimioterapia, terapia hormonal e terapia biológica. |
| **Estágio IIIb** | O tumor infiltra a parede torácica ou causa inchaço ou ulceração da mama ou é diagnosticado como câncer de mama inflamatório. Pode ou não ter se espalhado para os linfonodos axilares, mas não atinge outros órgãos do corpo. | |
| **Estágio IV** | É o câncer metastático. O tumor de qualquer tamanho espalhou-se para outros locais do corpo como ossos, pulmões, fígado, rins, intestinos, cérebro. | O tratamento pode envolver cirurgia, radioterapia, quimioterapia, terapia hormonal ou uma combinação desses tratamentos. |

Os trabalhos relacionados na Tabela 2 motivam a realização de uma investigação de métodos para diagnósticos auxiliados por computador, com o intuito de indicar possíveis regiões de interesse em imagens mamográficas ao radiologista e classificá-las em tecidos normais, benignos ou malignos.

Atualmente poucos sistemas CADs, que fornecem uma sugestão de diagnóstico, foram homologados pela agência americana de administração de alimentos e medicamentos (FDA, do inglês, *food and drug administration)* (CALAS; GUTFILEN; PEREIRA, 2012). No Brasil, não foi encontrado nenhum sistema comercial que esteja em ação dentro dos hospitais ou algum sistema que seja homologado pelo Ministério da Saúde (MS). Com isto, faz-se necessária a construção de um sistema de sugestão de diagnóstico para que se possam realizar análises nos hospitais públicos, visando uma futura homologação, tanto pelo MS quanto pelo FDA.

Com um sistema CAD, além do aumento da sensibilidade e especificidade, os hospitais poderão ganhar em média 19% do tempo da análise de uma mamografia (TCHOU et al., 2010). Será obtida também uma economia em torno de R$ 600,00 por mamografia diagnosticada como falso positivo, custo referente à biópsia não realizada. Com tal conduta, o único custo efetivo será a mamografia que tem o valor aproximado de R$ 100,00. (CONSELHO REGIONAL DE MEDICINA DO ESTADO DO MATO GROSSO (CRM-MT), 2012).

Haverá também a melhora do bem estar das pacientes, já que 29% das mulheres com diagnóstico falso positivo relataram ansiedade sobre o câncer de mama comparada com as 13% com a mamografia negativa (GRAM; LUND; SLENKER, 1990). Após a biópsia de pacientes diagnosticadas com um resultado falso positivo, 27% das pacientes queixaram-se de dor no peito e 33% tiveram sua sensibilidade sexual reduzida (GRAM; LUND; SLENKER, 1990).

**Tabela 2: Sistemas de diagnóstico de câncer de mama auxiliado por computador. A primeira coluna é a referência do trabalho, a segunda é a metodologia utilizada, a terceira é o banco de imagens e a quarta é a quantidade de imagens utilizadas. As últimas colunas são os resultados apresentados, sendo que FLL é a fração de lesão localizada, FNL é a fração de não lesão localizada, ACC é a acurácia, SEN é a sensibilidade e ESP é a especificidade.**

| Trabalho | Metodologia | Banco de imagens | Imagens | Segmentação | | Classificação (%) | | |
|---|---|---|---|---|---|---|---|---|
| | | | | FLL | FNL | ACC | SEN | ESP |
| WEI et al., 1995 | Análise de textura em multiresolução e transformada de *wavelet.* | *Department of Radiology at the University of Michigan* | 672 | - | - | 86 | - | - |
| CAMPANINI et al., 2002 | Transformada de *wavelet* e Máquina de Vetor de Suporte. | DDSM | - | 0,83 | 3,1 | - | 84 | - |
| CHRISTOYIANNI et al., 2002 | Análise de componentes independentes e Redes Neurais de Função Base Radial. | MIAS | 119 | - | - | 88,3 | - | - |
| ZHANG; VERMA; KUMAR, 2004 | Algoritmos Genéticos e Redes Neurais. | DDSM | 117 | - | - | 87,2 | - | - |
| CHEN; CHANG, 2004 | Unidade de codificação de textura e redes neurais probabilísticas. | MIAS | 59 | - | - | 71 | - | - |
| LIM; ER, 2004 | Rede neuro-nebulosa genérica e dinâmica. | DDSM | 343 | - | - | 70 | 95 | 60 |
| OZEKES; OSMAN; ÇAMURCU, 2005 | *Mass template.* | MIAS | 52 | - | 0,33 | 81 | - | - |
| CAMPOS; SILVA; BARROS, 2005 | Análise de componentes independentes e redes neurais probabilísticas. | MIAS | 200 | - | - | 97,3 | 96 | 100 |
| MARTINS et al., 2006 | Matriz de co-ocorrência e redes neurais Bayesianas. | MIAS | 218 | - | - | 86,9 | 83,4 | 95 |
| MASOTTI, 2006 | Análise de componentes principais. | DDSM | 588 | - | - | - | 90 | - |
| MASALA, 2006 | Transformada de *wavelet* e redes neurais artificiais. | MAGIC-5 | - | 0,88 | 2,22 | - | - | - |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| MOAYEDI et al., 2007 | Máquinas de vetor de suporte e rede neuro-nebulosa. | MIAS | 60 | - | - | 91,52 | - | - |
| BRAZ et al., 2007 | Características geoestatísticas e transformada de *wavelet.* | DDSM | 2048 | - | - | 98,24 | 98,1 | 98,3 |
| TIMP; VARELA; KARSSEMEIJER, 2007 | Análise de mudança temporal. | *Dutch Breast Cancer Screening Programme* | 465 | - | - | 79 | 88 | - |
| YUAN et al., 2007 | Modelo de contorno geométrico ativo. | FFDM | - | 0,85 | - | - | - | - |
| ELTONSY; TOURASSI; ELMAGHRABY, 2007 | Modelo morfológico concêntrico. | DDSM | 270 | - | 0,6 | - | 81 | - |
| (KOM; TIEDEU; KOM, 2007) | Limiar Local Adaptativo. | *Yaounde Gynaeco–Obstetric and Pediatric Hospital* (YGOPH). | 49 | 0,95 | - | - | - | - |
| WANG; GAO; LI, 2007 | Comentários relevantes e máquinas de vetor de suporte. | DDSM | 350 | - | 3,6 | 90,6 | - | - |
| MARTINS et al., 2007 | Neural GAS e Função K de Ripley. | DDSM | 997 | - | 0,93 | 89,3 | - | - |
| ZHANG et al, 2008 | Rede neural de retropropagação. | FFDM | 30 | 0,82 | - | 83,3 | - | - |
| DOMÍNGUEZ; NANDI, 2008 | Realce estatístico, segmentação limiar de multinível e seleção de região. | MIAS | 57 | 0,8 | 2,3 | - | - | - |
| NUNES; SILVA; PAIVA, 2010 | Índice de diversidade de Simpson e máquinas de vetor de suporte. | DDSM | - | - | 0,55 | 83,94 | 83,24 | 84,14 |
| SAMPAIO et al., 2011 | Redes neurais celulares, funções geoestatísticas e máquinas de vetor de suporte. | DDSM | 623 | - | 0,84 | 84,62 | 80 | 85,68 |
| XU; PEI, 2011 | Combinação hierárquica. | DDSM | 368 | - | 5,2 | 96,2 | - | - |
| ISA; SIONG, 2012 | Região de restrição crescente. | MIAS | 322 | - | 3,9 | - | 94,59 | |

## 1.2. Objetivo

### 1.2.1. Objetivo Geral

Propor um sistema de diagnóstico auxiliado por computador, baseado em métodos estatísticos, que detecte regiões de interesse em mamografias digitalizadas e as classifique como tecido normal, benigno ou maligno.

### 1.2.2. Objetivos Específicos

- Avaliar a técnica de codificação eficiente (análise de componentes independentes), como descritor de anormalidades do tecido mamário, comparando-a com outras técnicas de extração de características (*wavelets* de Gabor e análise de componentes principais);
- Segmentar regiões de interesse em mamografias utilizando as funções bases das técnicas de extração de características como um banco de filtros de imagens auxiliado por algoritmos de agrupamento;
- Comparar a efetividade dos algoritmos de agrupamentos conhecidos por *k-means*, nebuloso *c-means* e mapa auto-organizável;
- Classificar as regiões de interesse em massa ou não-massa e em benigno ou maligno, quando for o caso, utilizando as técnicas de extração de características e o classificador análise discriminante linear (LDA, do inglês *linear discriminat analysis)*;
- Desenvolver e registrar o software de diagnóstico precoce de câncer de mama a partir de imagens mamográficas, podendo ser utilizado por qualquer profissional.

### 1.3. Organização do trabalho

Neste capítulo, foi descrita a relevância da detecção precoce do câncer de mama e do papel da mamografia na detecção precoce de lesões na mama. Também foram ressaltados os benefícios dos sistemas de diagnóstico auxiliado por computador e os objetivos deste trabalho.

No capítulo 2, são descritas algumas características dos mamógrafos, das imagens mamográficas e do câncer de mama. Estes fundamentos serão importantes para melhor entendimento deste trabalho.

O capítulo 3 descreve o conceito de processamento digital de imagens e a importância da elaboração de um Sistema de Visão Artificial no processamento de uma imagem. Também são abordadas as técnicas utilizadas para extração de características, *wavelets* de Gabor, análise de componentes principais e análise de componentes independentes. Todas estas técnicas serão utilizadas tanto para a segmentação das regiões de interesse em mamografias quanto na classificação destas regiões em regiões de massa ou não-massa. Os algoritmos de agrupamentos utilizados neste trabalho, *k-means*, nebulo *c-means* e mapa auto-organizável, para realizar a segmentação das imagens e do algoritmo de análise discriminante linear utilizado para classificar estas regiões de interesse, também serão descritos.

O capítulo 4 trata exclusivamente da metodologia da segmentação de regiões de interesse em imagens mamográficas. Este método é dividido em três etapas: aquisição de imagens, extração de características e agrupamento. Em seguida serão mostrados os resultados obtidos por esta metodologia e as discussões e conclusões.

No capítulo 5 são abordados a metodologia e os resultados da classificação de regiões de interesse. Esta metodologia é dividida em três passos: aquisição de imagens, extração de características e classificação. Finalizando o capítulo com as discussões e conclusões.

O capítulo 6 finaliza este trabalho com as considerações finais e perspectivas futuras.

## 2. MAMOGRAFIA E O CÂNCER DE MAMA

Em 1998, o Ministério da Saúde (MS) publicou no Diário Oficial da União as "Diretrizes de proteção radiológica em radiodiagnóstico médico e odontológico" (MINISTÉRIO DA SAÚDE. SECRETARIA DE VIGILÂNCIA SANITÁRIA, 1998) regulamentando entre outras coisas as especificações e os requisitos técnicos mínimos para o mamógrafo, que deve ter um gerador trifásico ou de alta frequência, tubo especificamente projetado para mamografia (com janela de berílio), filtro de molibdênio, escala de tensão em incrementos de 1 KV, dispositivo de compressão firme (força de compressão entre 11 a 18 quilograma-força), diafragma regulável com localização luminosa, distância foco-filme não inferior a 30 centímetros e tamanho de ponto focal não superior a 4 milímetros.

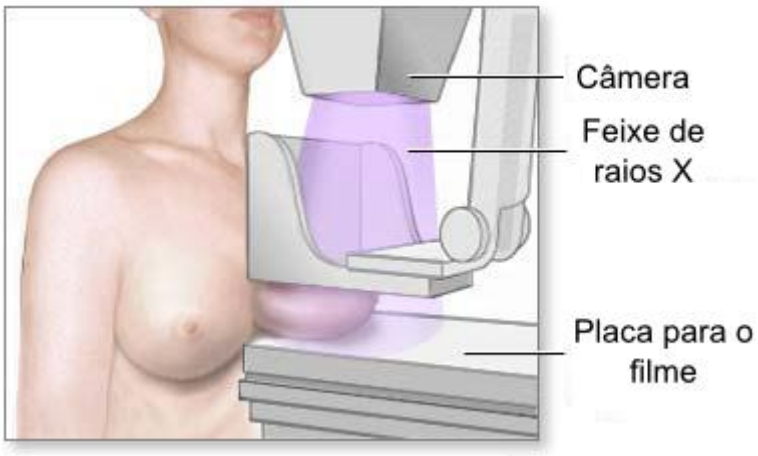


**Figura 1: Ilustração de um mamógrafo, onde se pode ver a câmera, o feixe de raios-X e a placa para o filme onde fica armazenada a imagem mamográfica. Na mamografia, cada mama é comprimida horizontalmente e, em seguida, obliquamente, armazenando uma imagem de raios-X de cada posição. Adaptada de (BRANDÃO, 2012).**

O reconhecimento de estruturas que possam indicar a presença de câncer ocorre 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), conforme Figura 2(a). Geralmente, microcalcificações e massas aparecem em tonalidades mais claras na imagem obtida após a revelação do filme mamográfico, observe Figura 2(b), mas esta diferenciação fica prejudicada

em imagens de mamas densas. Por esse motivo, muitas vezes a descoberta do câncer de mama em mulheres com menos de 40 anos de idade acontece quando o tumor já apresenta um desenvolvimento avançando, o que dificulta o tratamento da doença. O diagnóstico dos cânceres não palpáveis só é possível através da realização de mamografias minuciosas, em que cada detalhe é de extrema importância para evitar os diagnósticos falsos positivos e falsos negativos (BLAND; COPELAND, 1994, 2000).

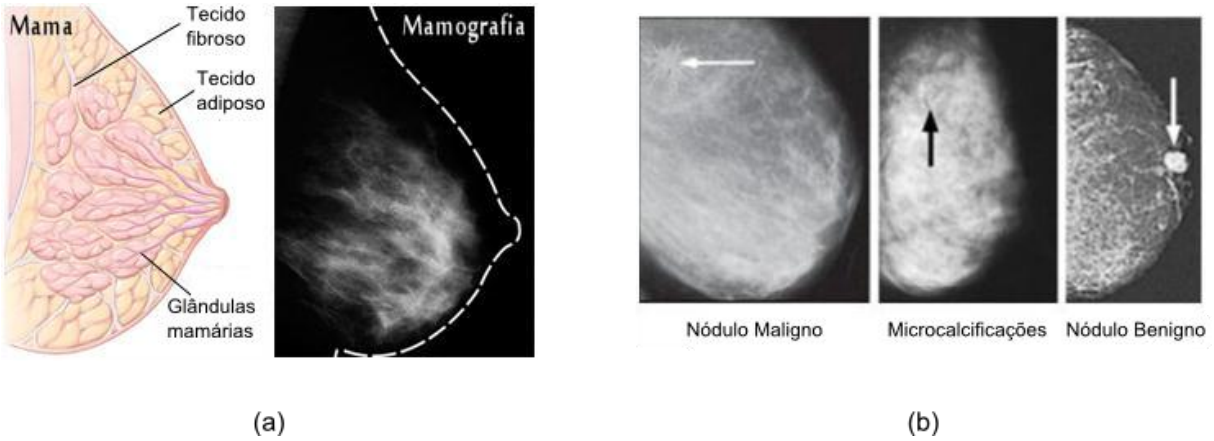


**Figura 2: (a) É a ilustração de uma mama com a sua representação mamográfica, onde se observa a apresentação destes tecidos na mamografia. Em (b) é ilustrado um exemplo de nódulo maligno, microcalcificações e nódulo benigno, respectivamente. Adaptada de (HARVARD MEDICAL SCHOOL - PORTUGAL PROGRAM, 2012).**

A mamografia é um exame que utiliza baixa tensão (KV) e altas correntes (mA) para gerar alto contraste, necessário na identificação das estruturas que compõem a mama, todas com densidade semelhante.

Na realização da mamografia, deve-se utilizar compressão eficiente, entre 13 a 15 quilogramas-força, para obtenção do exame (na prática, em aparelhos que não indicam automaticamente a força de compressão utilizada, pode-se comprimir até a pele ficar tensa e/ou até o limite suportado pela paciente).

Segundo o INCA (MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2007), as vantagens da compressão são as seguintes:

- Reduz a dose de radiação, porque diminui a espessura da mama;

- Aumenta o contraste da imagem, porque a redução da espessura da mama diminui a dispersão da radiação;
- Aumenta a resolução da imagem, porque restringe os movimentos da paciente;
- Diminui distorções, porque aproxima a mama do filme;
- "Separa" as estruturas da mama, diminuindo a superposição e permitindo que lesões suspeitas sejam detectadas com mais facilidade e segurança;
- Diminui a variação na densidade radiográfica ao produzir uniformidade na espessura da mama.

As incidências seguem padronização, tanto do posicionamento da paciente quanto da angulação do tubo. Na mamografia, são utilizadas as incidências básicas e as incidências complementares. As incidências básicas, craniocaudal (CC) ilustrada na Figura 3 e médio-lateral oblíqua (MLO) ilustrada na Figura 4, representam a base de todos os exames e são indispensáveis. As incidências complementares esclarecem situações detectadas nas incidências básicas, servem para realizar manobras e estudar regiões específicas.

A incidência médio-lateral-oblíqua é a mais eficaz, pois mostra uma quantidade maior de tecido mamário e inclui estruturas mais profundas do quadrante superior externo e do prolongamento axilar. A incidência craniocaudal tem como objetivo incluir todo o material póstero-medial, completando a médio-lateral-oblíqua, que com frequência não está totalmente demonstrado na incidência MLO. Esta técnica permite também mais compressão da mama, uma vez que não inclui a axila, resultando em uma definição superior da arquitetura mamária e das lesões. Os radiologistas estudam as incidências craniocaudais e as médio-laterais aos pares de modo a permitir a comparação de regiões simétricas. Qualquer assimetria pode ser indício de patologia.

A mamografia deve ser realizada seguindo os critérios do documento de consenso de 2004 (MINISTÉRIO DA SAÚDE. SECRETARIA DE ATENÇÃO A SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2004b), onde o Ministério da Saúde recomenda que o exame mamográfico seja realizado com um intervalo máximo de dois anos entre os exames para as mulheres entre 50 a 69 anos; e anualmente, a partir dos 35 anos, para as mulheres pertencentes a grupos populacionais com risco elevado de desenvolver câncer de mama.

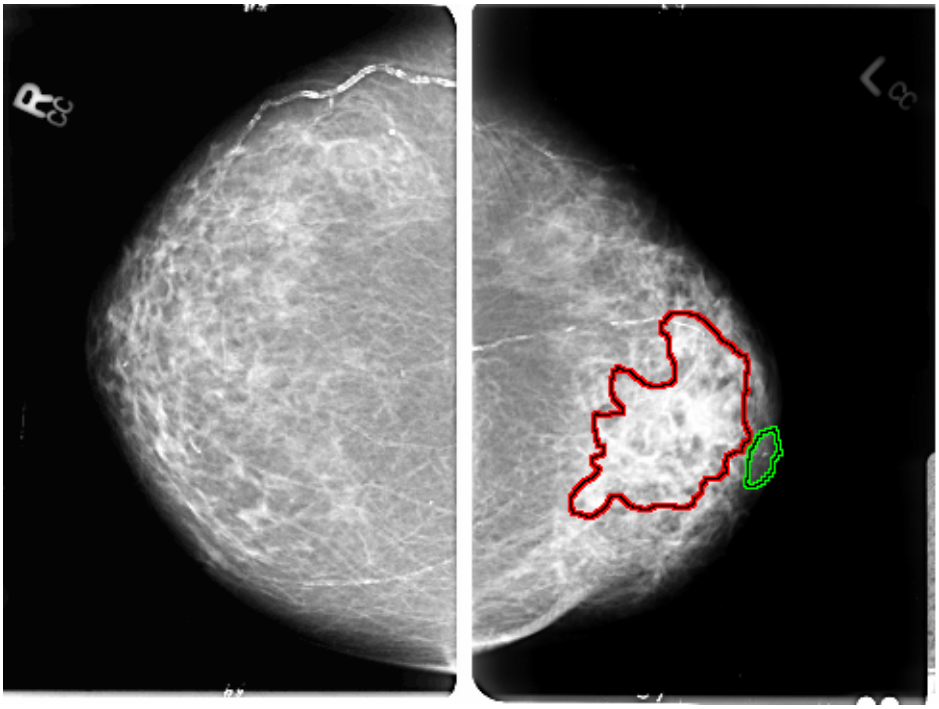

**Figura 3: Mamografia com incidência craniocaudal. Dois tumores malignos devidamente marcados na mama esquerda de uma paciente de 65 anos. Fonte: banco de dados DDSM (HEATH et al., 1998). Identificação Volume: câncer 01 Caso: B-3027-1.**

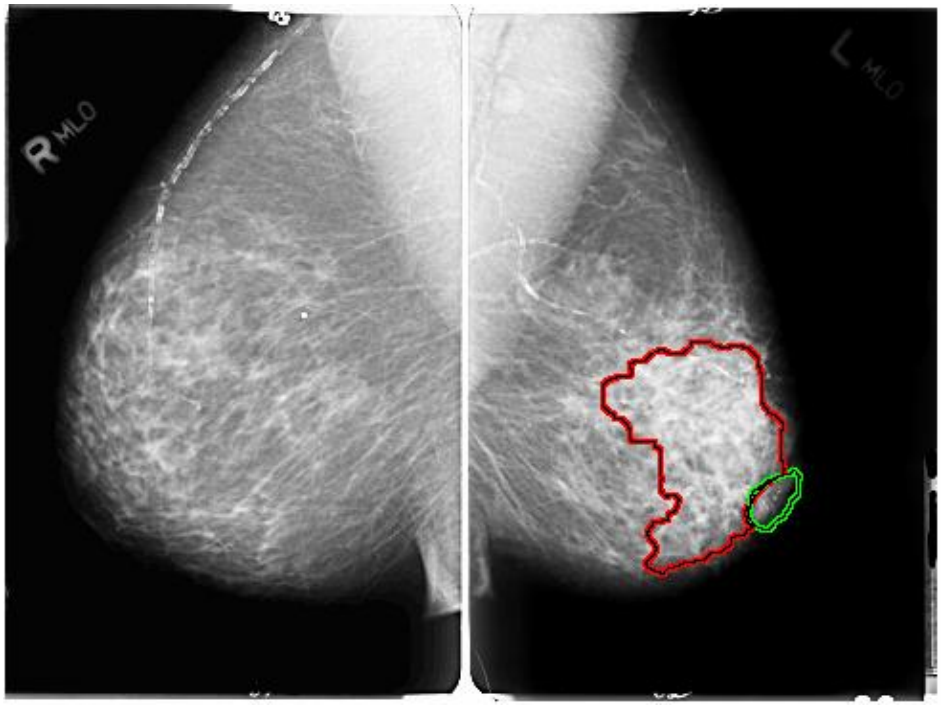

**Figura 4: Mamografia com a incidência médio-lateral oblíqua. Dois tumores malignos devidamente marcados na mama esquerda de uma paciente de 65 anos. Fonte: banco de dados DDSM (HEATH et al., 1998). Identificação Volume: câncer 01 Caso: B-3027-1.**

Segundo o INCA (MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2007), os nódulos devem ser analisados de acordo com o tamanho, contorno, limites e densidade.

- Tamanho: no caso das lesões não palpáveis este parâmetro é de importância relativa, pois os nódulos diagnosticados apenas pela mamografia apresentam pequenas dimensões. No caso dos nódulos ovalados, pode-se utilizar como medida o maior eixo; no caso dos nódulos arredondados, a medida representa o diâmetro;
- Contorno: os nódulos podem apresentar contorno regular, lobulado, microlobulado, irregular e espiculado. A suspeita de malignidade aumenta em função da ordem citada acima;

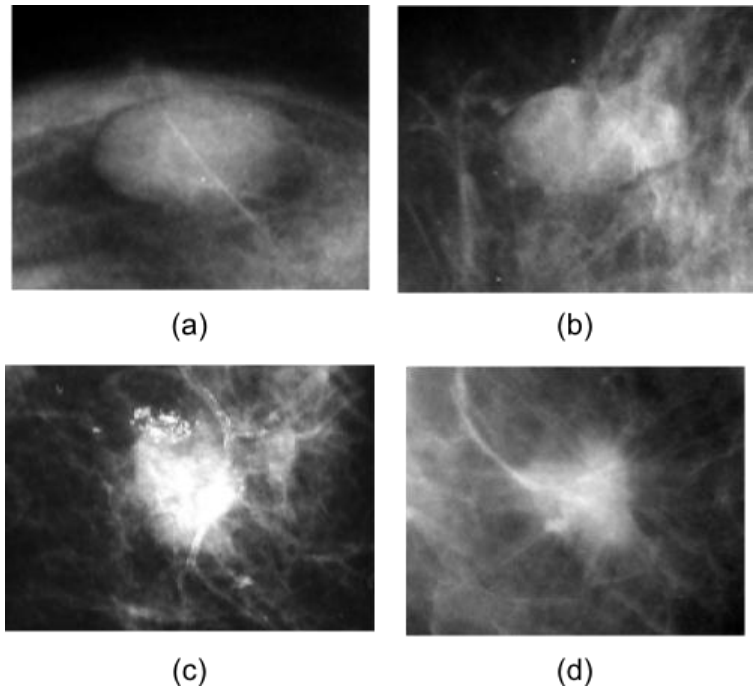


**Figura 5: Contorno dos nódulos. (a) Contorno regular; (b) contorno lobulado; (c) contorno irregular; (d) contorno espiculado.**

- Limites: os limites representam a relação do nódulo com as estruturas vizinhas e podem ser definidos, parcialmente definidos ou pouco definidos, quando a relação com as estruturas vizinhas é identificada em mais de 75%, entre 25% a 75% e menos do que 25% do contorno do nódulo, respectivamente. Limites mal definidos são mais sugestivos para malignidade do que limites parcialmente definidos e limites definidos;

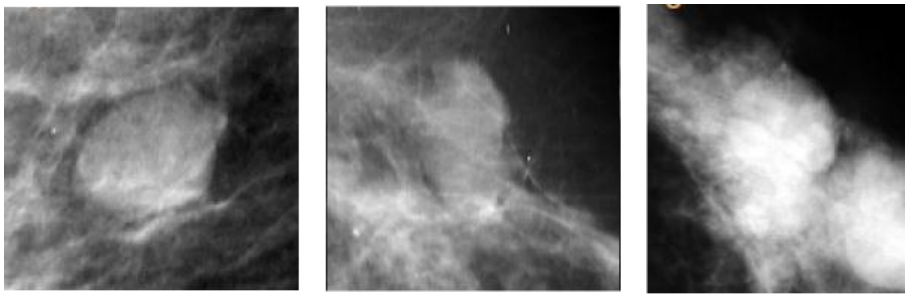


**Figura 6: Limite dos nódulos. (a) Limite definido; (b) limite parcialmente definido; (c) limite pouco definido.**

- Densidade: os nódulos podem ser densos, isodensos ao parênquima mamário, com baixa densidade, com densidade de gordura e com densidade heterogênea. Nódulos malignos geralmente têm densidade elevada, linfonodos intramamários têm densidade baixa, lipomas e cistos oleosos têm densidade de gordura e fibroadenolipomas têm densidade heterogênea.

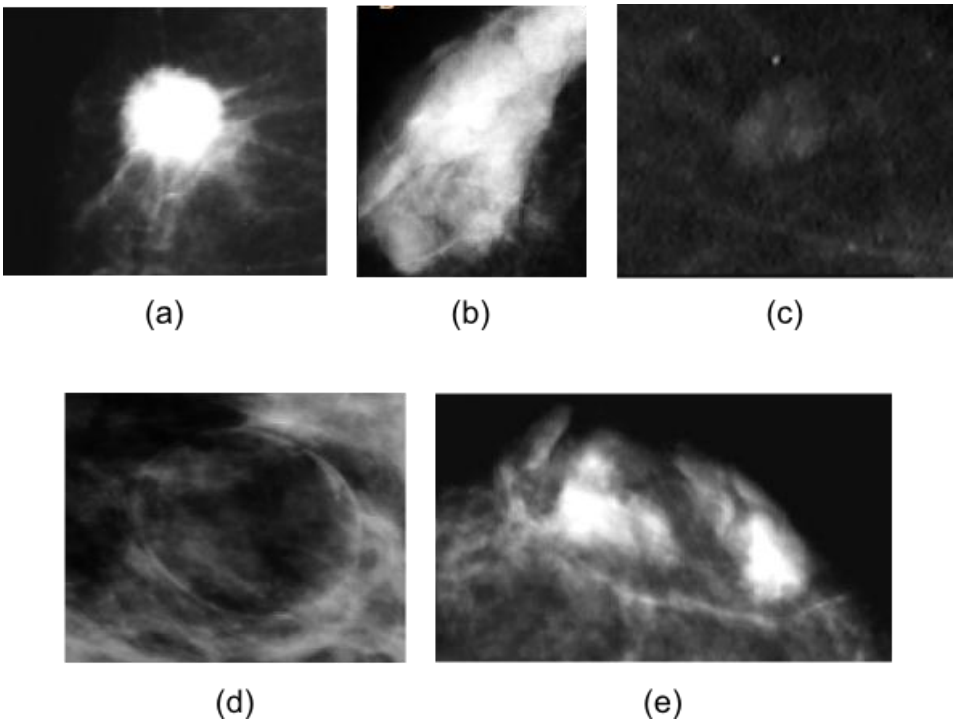


**Figura 7: Densidade dos nódulos. (a) Nódulo denso; (b) nódulo isodenso. (c) nódulo com baixa densidade; (d) nódulo com densidade de gordura; (e) nódulo com densidade heterogênea.**

O primeiro sinal de um câncer de mama é normalmente uma anormalidade detectada na mamografia antes que possa ser sentida pela própria mulher ou por um agente de saúde.

Grandes tumores podem ser evidenciados como um nódulo indolor. Sintomas como mudanças persistentes na mama, como espessamento cutâneo, retração cutânea, retração do complexo aréolo-papilar, corpo mamário com densidade difusamente aumentada e aspecto infiltrado, linfonodos axilares aumentados, densos e confluentes são menos comuns. Tipicamente, dores mamárias vêm de condições benignas e isto não é um sintoma inicial do câncer de mama (*AMERICAN CANCER SOCIETY*, 2007; MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2007).

Existe uma grande quantidade de resultados errados e divergentes em mamografias, principalmente se o radiologista for inexperiente (BARLOW et al., 2004). De acordo com o INCA (MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2007), a mamografia tem sensibilidade entre 88% a 93,1% e especificidade entre 85% a 94,2%. A precisão no diagnóstico está diretamente relacionada à idade da mulher, sendo muito menor nas mulheres jovens, que apresentam uma alta densidade de tecido mamário, devido à predominância de tecidos fibroglandulares na sua composição. Para garantir o desempenho da mamografia, a imagem obtida deve ter alta qualidade e, para tanto, são necessários: equipamento adequado, técnica radiológica correta, conhecimento, prática e dedicação dos profissionais envolvidos (MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER, 2007).

# 3. FUNDAMENTOS TEÓRICOS

Neste capítulo são abordadas as ferramentas utilizadas neste trabalho. Inicia-se com uma introdução a processamento digital de imagens, onde será passada a ideia de extração de característica e reconhecimento. As técnicas de extração de características abordada neste capítulo são as *wavelets* de Gabor, análise de componentes principais e análise de componentes independentes. Os algoritmos de agrupamentos citados no capítulo são o *k-means*, o nebuloso *c-means* e o mapa auto-organizável.

## 3.1. Processamento digital de imagens

O termo processamento digital de imagens geralmente se refere ao processamento de uma imagem por um computador. Isto implica em um processamento digital de dados bidimensionais. Uma imagem digital é um vetor de números reais ou complexos representados por um número finito de bits (JAIN, 1989).

O processamento de imagens digitais abrange uma ampla escala de hardware, software e fundamentos teóricos (GONZALES; WOODS, 2010). A seguir serão discutidos os passos fundamentais para a elaboração de um Sistema de Visão Artificial (SVA) capaz de adquirir, processar e interpretar imagens correspondentes a cenas reais (FILHO; NETO, 1999). A Figura 8 ilustra o fluxograma de um SVA.

Antes do primeiro passo é necessário ter o domínio do problema e do resultado, para que se possa preparar cada etapa com foco na solução. Neste trabalho, o problema é detectar e diagnosticar ROIs e o resultado seria a classificação dessas regiões de interesse em benigno, maligno ou normal.

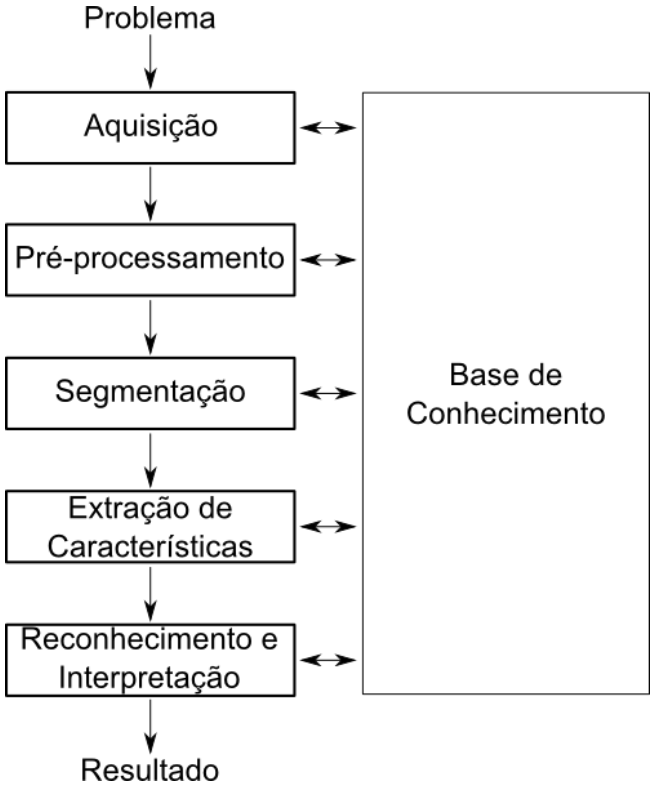


**Figura 8: Sistema Visual Artificial. As principais etapas de um SVA podem ser divididas em cinco passos: aquisição, pré-processamento, segmentação, extração de características e reconhecimento e interpretação. Todas as etapas estão interligadas através de uma base de conhecimento. Adaptada de (FILHO; NETO, 1999).**

O primeiro passo no processo é a aquisição da imagem, isto é, adquirir uma imagem digital. Para fazermos isto são necessários um sensor e um digitalizador. O sensor converterá a informação óptica em sinal elétrico e o digitalizador transformará a imagem analógica em imagem digital. Este sensor pode ser uma câmera ou um scanner. O tipo de sensor a ser utilizado varia conforme a aplicação.

O passo seguinte é o pré-processamento. Esta é a fase de aprimoramento da qualidade da imagem para que as chances de sucesso dos processos seguintes sejam maiores. Como exemplo de pré-processamento tem-se o realce de contraste através da equalização do histograma (Figura 9), filtragens para retirada de ruídos (Figura 10) e a aplicação de operadores morfológicos, como o de dilatação e o de erosão para evidenciar características (Figura 11).

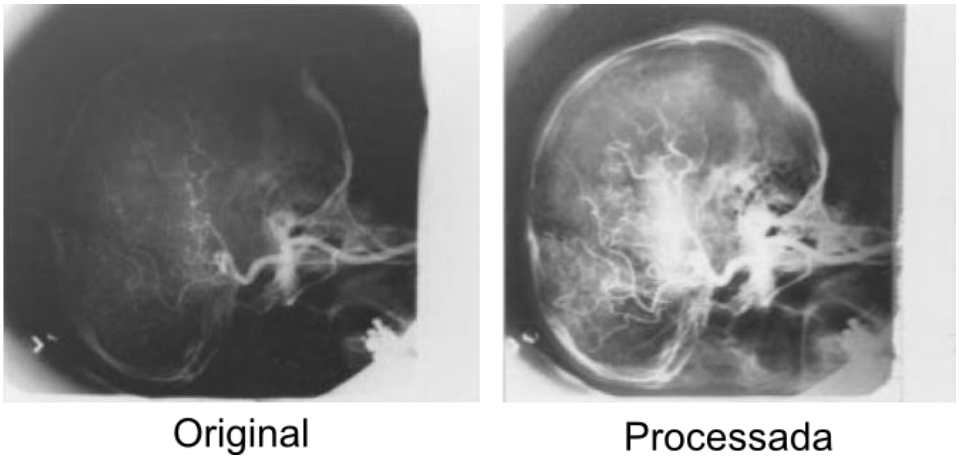


**Figura 9: Equalização do histograma. Exemplo de uma equalização de histograma em uma imagem médica de escaneamento cerebral. Observa-se que houve um realce na imagem. Adaptada de (PRATT, 2001).**

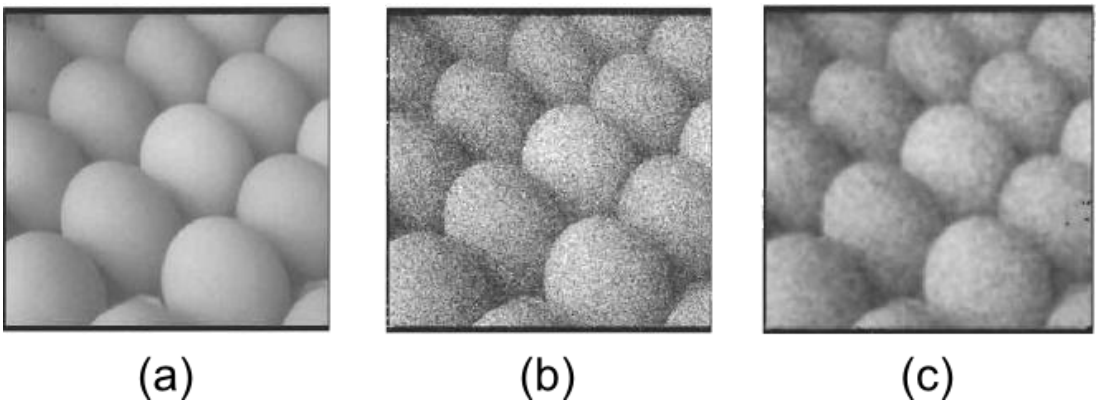


**Figura 10: Exemplo da aplicação de um filtro de média. (a) É a imagem original. Em (b) foi adicionado um ruído gaussiano à imagem original. Em (c) observa-se o resultado da filtragem utilizando uma máscara de média 5x5. Verifica-se que houve uma redução destes ruídos. Adaptada de (BOVIC; ACTON, 2001).**

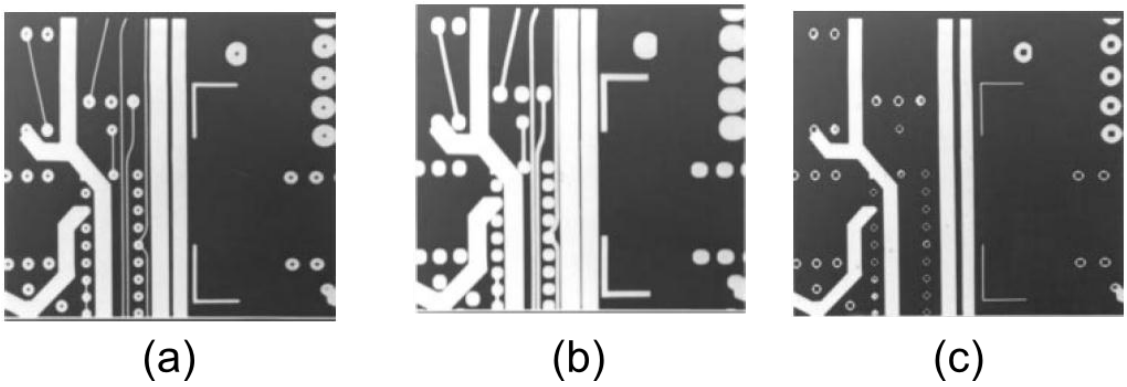


**Figura 11: Exemplo de aplicação de operadores morfológicos. Em (a) observa-se a imagem original. (b) É o resultado da operação de dilatação e (c) é o resultado da operação de erosão da imagem original. Constata-se que houve uma maior evidência em algumas áreas, dependendo do filtro aplicado. Adaptada de (PRATT, 2001).**

O terceiro passo é o da segmentação. Definida em termos gerais, a segmentação divide uma imagem de entrada em partes ou objetos constituintes. Em geral, a segmentação automática é uma das tarefas mais difíceis no processamento de imagens digitais. Por um lado, um procedimento de segmentação robusto favorece substancialmente a solução bem sucedida de um problema de imageamento. Por outro lado, algoritmos de segmentação fracos ou erráticos quase sempre asseveram falha no processamento. No caso da classificação de nódulos mamários, a segmentação deve extrair as regiões de interesse, isto é, possíveis regiões com nódulos, e descartar todo o resto da mamografia.

O processo de descrição, também chamado seleção de característica, procura extrair características que resultem em alguma informação quantitativa de interesse ou que sejam básicas para discriminação entre classes de objetos. No caso das imagens mamográficas, a forma e a textura são características poderosas que auxiliam na diferenciação entre nódulos benignos ou malignos e até mesmo no caso de tecidos saudáveis.

O último estágio envolve o reconhecimento e interpretação. Reconhecimento é o processo que atribui um rótulo a um objeto, baseado na informação fornecida pelo seu descritor (GONZALES; WOODS; EDDINS, 2009; GONZALES; WOODS, 2010). Neste trabalho, a identificação de um nódulo, por exemplo, benigno requer a associação dos descritores para aquele nódulo com o rótulo benigno e assim ocorre também para as classes malignas e normais.

Todas as tarefas das etapas descritas acima pressupõem a existência de um conhecimento sobre o problema a ser resolvido, armazenado em uma base de conhecimento, cujo tamanho e complexidade podem variar. Idealmente, esta base de conhecimento deveria não somente guiar o funcionamento de cada etapa, mas também permitir a realimentação entre elas.

### 3.2. Extração de características

Supõe-se que uma das estratégias utilizadas pelo cérebro para representar informação seja o princípio da codificação eficiente (BARLOW, 1961). Este conceito foi proposto por Horace Barlow, em 1961, como um modelo teórico para a codificação das informações sensoriais pelo sistema nervoso. Para ele, um modelo eficiente seria aquele que minimizasse a quantidade de impulsos nervosos utilizados para transmitir a informação desejada. Para SMITH e LEWICKI (2006), os sistemas sensoriais avançaram de tal forma que as estratégias de codificação eficiente transmitem a máxima informação para o cérebro enquanto minimizam a energia necessária e recursos neurais.

No desenvolvimento de sua teoria, Barlow foi inspirado por conceitos da teoria da informação. Ele definiu que os caminhos neurais percorridos por informações sensoriais são similares a canais de comunicação. Através de conceitos como capacidade de canal e redundância, Barlow sugeriu que a codificação neural é realizada de forma a maximizar a capacidade de canal e reduzir a redundância na informação transmitida.

A teoria da informação desempenha um papel natural em modelos de sistemas neurais, fornecendo uma definição quantitativa única para informação (COVER; THOMAS, 1991). Barlow reconheceu a importância da teoria da informação neste contexto, e levantou a hipótese de que a codificação eficiente da informação visual poderia servir como uma restrição fundamental no processamento neural (BARLOW, 1961). Isto é, um grupo de neurônios deve codificar a informação o mais compacta possível, a fim de uma utilização mais eficaz dos recursos computacionais disponíveis. Matematicamente, isto é expresso como um desejo para maximizar a informação de que as respostas neuronais possam fornecer sobre o ambiente visual (SIMONCELLI, 2003).

A forma mais simples desta hipótese (em particular, ignorando o ruído em respostas neurais) dissocia-se naturalmente em duas declarações separadas: uma sobre as estatísticas individuais de respostas neurais e um segundo em relação às estatísticas conjuntas da resposta de uma população (PANZERI et al, 1999; SIMONCELLI; OLSHAUSEN, 2001; DAYAN; ABBOTT, 2001).

- A resposta de um neurônio individual para o ambiente natural deve utilizar totalmente a sua capacidade de saída, dentro dos limites de quaisquer restrições sobre a resposta (SIMONCELLI, 2003).
- As respostas de neurônios diferentes para o ambiente natural devem ser estatisticamente independentes uns dos outros. Em outras palavras, a informação transportada por cada neurônio não deve ser redundante com a que é transportada pelos outros. Isto também é consistente com a noção de que o sistema visual se esforça para decompor uma cena em componentes estatisticamente independentes (SIMONCELLI, 2003).

Um exemplo prático da codificação eficiente encontra-se no sistema visual humano, que é composto por cerca de seis milhões de cones fotorreceptores nos olhos, cada um fornece uma amostra de ponto da imagem da retina. Mas o nervo óptico, através do qual todos os sinais são transmitidos para o cérebro, contêm apenas 1,25 milhões de fibras nervosas, cada um dos quais podem levar menos informação que um único cone (LENNIE, 2003). Esta perda de informações no olho não é necessariamente problemática: a imagem da retina, como a maioria dos sinais sensoriais, é redundante, e uma parte da imagem pode, até certo ponto, ser prevista a partir da estrutura de peças vizinhas, mas para a predição adequada a informação correta tem de ser preservada (LENNIE, 2003).

O sistema visual, tal como outros sistemas sensoriais, necessita de filtros seletivos projetados para transmitir a estrutura importante de sinais naturais. Filtros são incorporados nos neurônios individuais na retina e estágios superiores da via visual, e tem propriedades características que determinam os atributos espaciais e temporais dos sinais que eles transmitem. A forma natural para representar os atributos espaciais do filtro neuronal é através de um mapa da região da retina a partir do qual o neurônio capta os sinais. Este mapa é chamado de campo receptivo do neurônio (LENNIE, 2003).

Os campos receptivos espaciais de células simples (V1) foram razoavelmente bem descritos fisiologicamente e podem ser caracterizados como sendo localizados, orientados, e filtrados por um filtro passa-faixa (HUBEL; WIESEL, 1968; DE VALOIS; ALBRECHT; THORELL, 1982; JONES; PALMER, 1987; PARKER; HAWKEN, 1988). Cada célula responde a estímulos visuais dentro de uma região restrita e contígua de espaço que está

organizado em subcampos excitatórios e inibitórios ao longo de uma direção particular, e a resposta de frequência espacial é geralmente um filtro passa-faixa com larguras de faixa na gama de 1-2 oitavas (OLSHAUSEN; FIELD, 1996a).

### 3.2.1. *Wavelets* Gabor

As células do córtex visual primário (V1) tem uma aparência ordenada, um mapa topográfico e um arranjo ordenado de dominância ocular e colunas de orientação (HUBEL; WIESEL, 1968; DE VALOIS; ALBRECHT; THORELL, 1982). Muitos neurônios são filtros especializados para características de estímulo, como a frequência espacial, orientação, cor, direção do movimento e disparidade (JONES; PALMER, 1987; PARKER; HAWKEN, 1988). E tem ainda emergido um modelo padrão razoavelmente bem acordado para V1 no qual as células simples calculam uma soma ponderada linear da entrada no espaço e no tempo (normalmente uma função do tipo Gabor) e a saída passa por uma não linearidade ponto-a-ponto, adicionalmente ao fato de ser sujeita a controle de ganho de contraste para evitar saturação da resposta, conforme Figura 12. Células complexas são igualmente explicadas em termos da soma dos resultados de uma associação local de células simples com propriedades semelhantes, mas em diferentes ajustes de posições ou fases. O resultado disto é que se deve pensar em V1 como um "banco de filtros de Gabor", apresentado na Figura 13. Existem muitos trabalhos que demonstram que este modelo básico se encaixa muito bem com os dados existentes, e muitos cientistas passaram a aceitar isso como um modelo de função de V1 (RINGACH, 2002; LENNIE, 2003; OLSHAUSEN; FIELD, 2004).

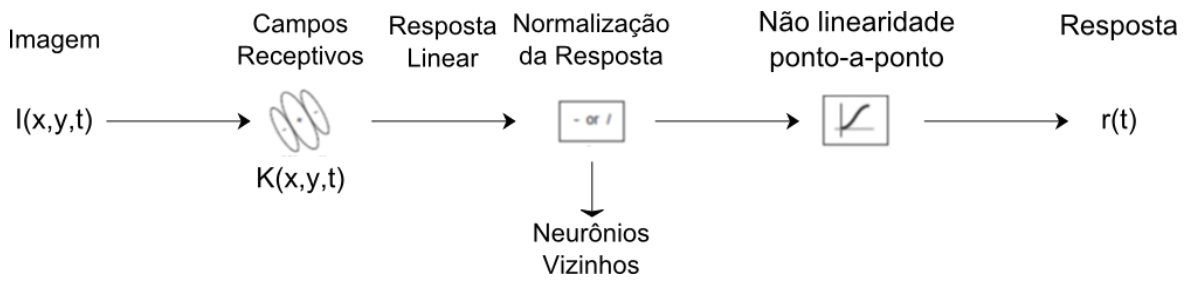


**Figura 12: Modelo padrão para respostas das células simples de V1. O neurônio calcula uma soma ponderada da imagem no espaço e no tempo. O resultado é normalizado pelas respostas de unidades vizinhas e passada através de uma não linearidade ponto-a-ponto (CARANDINI; HEEGER; MOVSHON, 1997). Adaptada de (OLSHAUSEN; FIELD, 2004).**

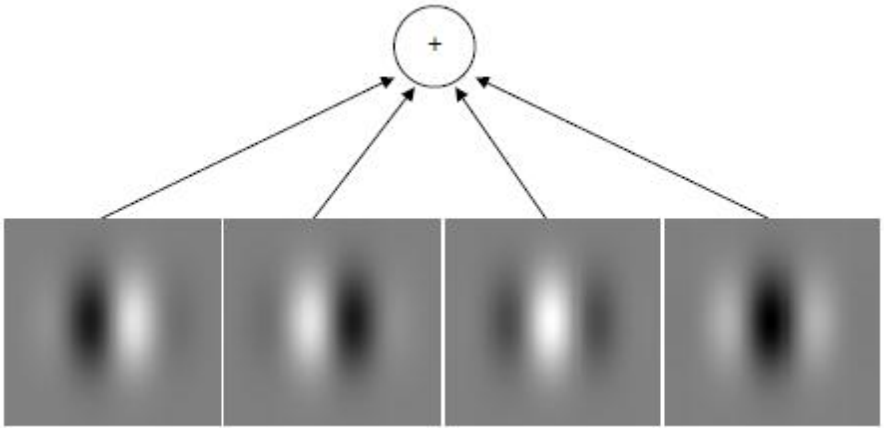

**Figura 13: Banco de filtros de wavelets de Gabor. Campos receptivos de uma célula complexa. Fonte: (LENNIE, 2003).**

Os filtros de Gabor são filtros passa-faixa com orientação sintonizável e larguras de faixa de frequências radiais. A transformada de Fourier dos filtros de Gabor são gaussianos deslocados em frequência. A representação de Gabor é provada ser ótima no sentido de minimizar a incerteza 2-D conjunta no espaço e frequência. Os filtros de Gabor têm formas semelhantes com a dos campos receptivos de células simples do córtex visual primário quando estimuladas por imagens naturais (OLSHAUSEN; FIELD, 1996b). Estes filtros são *kernels* de múltiplas escalas e orientações e são representados pela equação 1, que é a função 2-D de Gabor.

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)+2\pi j\omega x\right], \quad (1)$$

onde, $\sigma_x$ e $\sigma_y$ são os desvios padrões da elipse gaussiana ao longo de $x$ e $y$ e $\omega$ é a frequência do filtro espacial no domínio da frequência.

O banco de filtros de Gabor pode ser obtido por diferentes escalas e rotações de $g(x,y)$:

$$\begin{aligned} g_{mn}(x,y) &= a^{-m} g(x',y'), \\ x' &= a^{-m}(x\cos\theta + y\sin\theta), \\ y' &= a^{-m}(-x\sin\theta + y\cos\theta), \end{aligned} \quad (2)$$

onde $\theta = {}^{n}/_{\pi k}$ e $k$ é o número total de orientações, $m = 1, \cdots, M$, onde $M$ é o número de escalas.

### 3.2.2. Análise de Componentes Principais

A análise de componentes principais (PCA, do inglês *principal component analysis*) (JAIN, 1989; HYVÄRINEN; KARHUNEN; OJA, 2001; RENCHER, 2002) é uma técnica estatística poderosa que pode ser utilizada para estudar correlações entre dados, ou seja, determinar as suas direções principais. Entendem-se como direções principais o conjunto de vetores ortogonais sobre os quais os dados apresentam maior variância. O primeiro vetor representa a direção de máxima variância, o segundo vetor também está disposto segundo a direção de máxima variância, mas sob a condição de ser ortogonal ao primeiro, e assim sucessivamente para o restante dos vetores (Figura 14).

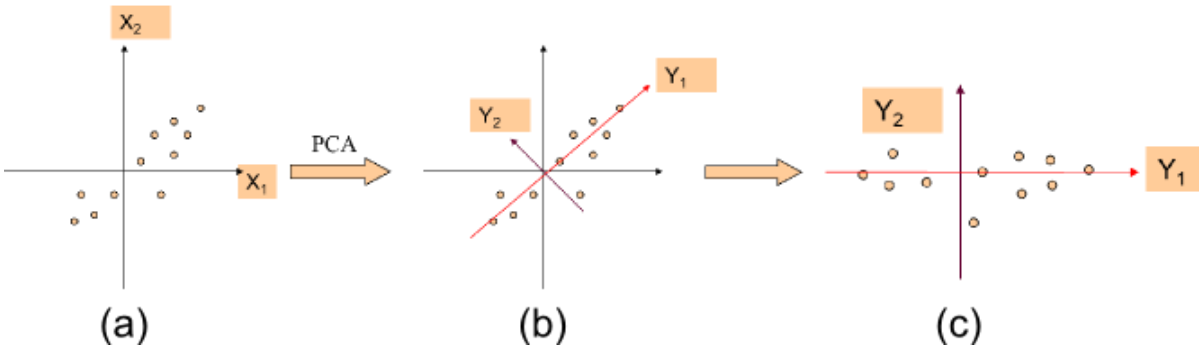


**Figura 14: Direção das componentes principais. (a) Representa o conjuto de dados originais. Após a PCA, em (b), é encontrado o vetor $Y_1$ que aponta na direção de maior variância dos dados e $Y_2$ que é o vetor que aponta para a segunda maior variância, obedecendo o critério de ser ortogonal ao vetor $Y_1$. Em (c) são ilustrados esses dados nestes novos eixos, descorrelacionando-os.**

Uma das principais aplicações da PCA é a redução de dimensionalidade através da eliminação das variáveis originais de menor variância. Embora a variabilidade total de um sistema seja definida por $n$ variáveis, geralmente muito desta variabilidade pode ser explicada por um número bem menor, $k$, de componentes principais. Desta forma, a quantidade de informação contida em $k$ é equivalente àquela existente nas $n$ variáveis originais.

Por isso, em muitas aplicações a PCA é utilizada como uma espécie de pré-processamento dos dados, servindo como entrada para outros modelos numéricos, tais como análise discriminante e máquinas de vetor de suporte. A vantagem, neste caso, está na redução

do número de parâmetros do modelo imediatamente seguinte à PCA, melhorando o desempenho e poupando tempo de processamento.

Análise de componentes principais tem sido muito utilizada em processamento de sinais biológicos, como a voz (ZOHARIAN; ROTHENBERG, 1981), ECG (CASTELLS et al., 2007) e EEG (JIN; WANG, X.; WANG, B., 2007). Como a PCA utiliza apenas estatística de segunda ordem, para obter um banco de filtros eficientes, ou seja, atingir o critério de independência estatística, os dados de onde serão retiradas as componentes principais devem ter uma distribuição gaussiana, caso contrário resultará em um código menos eficiente e insuficiente para explicar as propriedades do sinal de entrada.

#### 3.2.2.1. Cálculo das componentes principais

A redundância entre as variáveis é medida pela correlação entre suas observações. Retirando a média, a correlação e a covariância são iguais. Então, considerando que $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}$ é o conjunto de sinais de entrada, onde cada $\mathbf{x}$ é um processo aleatório de média zero, $\mathbf{C}_x$ é a autocovariância de $\mathbf{x}$ dada por:

$$\mathbf{C}_x = E\{\mathbf{X}\mathbf{X}^T\}. \tag{3}$$

As componentes principais de $\mathbf{X}$ estão contidas em uma matriz ortogonal $\mathbf{V}$ que realiza uma transformação linear de $\mathbf{X}$ em $\mathbf{Z}$, tal que a energia está maximamente concentrada, conforme equação 4:

$$\mathbf{Z} = \mathbf{V}^T\mathbf{X}, \tag{4}$$

onde, os elementos de $\mathbf{Z}$ são mutualmente descorrelacionados e a sua matriz de autocovariância deve, pois, ser diagonal, de forma que:

$$\mathbf{C}_z = \mathrm{E}\{\mathbf{Z}\mathbf{Z}^T\} = \mathbf{\Lambda}. \tag{5}$$

Sabe-se, porém que:

$$\mathbf{C}_z = E\{\mathbf{Z}\mathbf{Z}^T\} = E\{\mathbf{V}^T\mathbf{x}\mathbf{x}^T\mathbf{V}\} = \mathbf{V}^T \mathrm{E}\{\mathbf{x}\mathbf{x}^T\}\mathbf{V} = \mathbf{V}^T\mathbf{C}_x\mathbf{V}. \quad (6)$$

Das equações 5 e 6, tem-se que:

$$\mathbf{V}^T\,\mathbf{C}_x\mathbf{V} = \mathbf{\Lambda} \quad (7)$$

Portanto, $\mathbf{V}^T$ é a matriz ortogonal que diagonaliza a matriz $\mathbf{C}_x$. Como resultado clássico da álgebra, $\mathbf{V}^T$ é a matriz cujas linhas são os autovetores da matriz de $\mathbf{C}_x$, correspondentes aos autovalores em ordem crescente de variância. $\mathbf{\Lambda}$ é uma matriz diagonal cujos elementos são os autovalores de $\mathbf{C}_x$, ou correspondentemente, as variâncias de $\mathbf{Z}$, em ordem decrescente de energia.

Para se obter todo o conjunto de componentes principais, a equação de autovetores de $\mathbf{C}_x$ deve ser resolvida, como:

$$\begin{aligned} \mathbf{C}_x\mathbf{V} &= \mathbf{\Lambda}\mathbf{V}, \\ (\mathbf{C}_x - \mathbf{\Lambda}\mathbf{I})\mathbf{V} &= 0, \end{aligned} \quad (8)$$

onde,

$$|\mathbf{C}_x - \mathbf{\Lambda}\mathbf{I}| = \mathbf{0}. \quad (9)$$

### 3.2.3. Análise de Componentes Independentes

A fim de codificar os dados de forma mais eficiente, as respostas neuronais devem ser independentes umas das outras. Isto significa que a codificação da informação deve fazer uso igual de todas as possíveis combinações de padrões de ativação (SIMONCELLI; OLSHAUSEN, 2001). Uma independência estatística entre os dois sinais significa que o conhecimento de um sinal não fornece nenhuma informação sobre o outro sinal.

A análise de componentes independentes (ICA, do inglês *independent component analysis*) é um método visto como uma extensão da análise de componentes principais, já que para obter independência, garante-se a descorrelação, que é um resultado da PCA (HYVÄRINEN; KARHUNEN; OJA, 2001). A ICA foi desenvolvida no contexto de separação cega de fontes (BSS, *blind source separation*), em que o problema é definido na estimação da saída de uma fonte conhecida, quando esta fonte recebe vários sinais misturados e desconhecidos.

A ICA tem sido aplicada em diversas áreas, como por exemplo: processamento de imagens naturais (KARKLIN; SIMONCELLI, 2011), sons naturais (LEWICKI, 2002), fala (CAVALCANTE et al., 2006), compressão de imagens (SOUSA et. al., 2007), ECG (GUILHON; BARROS; COMANI, 2007), percepção visual (PENG-LU et al., 2011) e outras.

#### 3.2.3.1. Definições

Considere que sejam observadas $n$ misturas lineares $x_1, \cdots, x_n$, modeladas como uma combinação linear de $n$ funções bases, dadas por:

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \cdots + a_{in}s_n \,, \forall i = 1, \cdots, n \qquad (10)$$

e que cada mistura $x_i$, assim como cada componente independente $s_1, \cdots, s_n$ seja uma variável aleatória e $a_{ij}$ os coeficientes (pesos) da mistura linear. Assume-se que tanto as variáveis da mistura quanto aquelas das componentes independentes têm média zero. Por conveniência, será usada a notação vetorial em vez de somas, como aquelas vistas na equação 10, utilizando letras minúsculas e maiúsculas, para representar, respectivamente, vetores e matrizes. Dessa maneira, pode-se reescrever a equação anterior da seguinte forma:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \qquad (11)$$

O modelo estatístico definido na equação 11 é chamado de modelo de análise de componentes independentes. Este modelo descreve os dados observados pelo processo de

mistura das componentes independentes $\mathbf{s}_j$, que não podem ser observadas diretamente. É preciso estimar tanto $\boldsymbol{s}$ quanto a matriz de mistura **A**, que também é desconhecida, pois a única informação conhecida é o vetor **x**. Para tanto, é preciso fazer suposições tão gerais quanto possível (HYVÄRINEN; KARHUNEN; OJA, 2001). Portanto, supõe-se que:

- As componentes $s_1, \cdots, s_n$ são estatisticamente independentes;
- As componentes têm distribuições não gaussianas.

#### 3.2.3.2. Definição de independência

Sejam $y_1$ e $y_2$ duas variáveis aleatórias quaisquer, elas serão ditas independentes se a ocorrência ou não ocorrência de $y_1$ não influenciar na ocorrência ou não ocorrência de $y_2$, e vice-versa. Matematicamente, independência estatística é definida em termos da densidade de probabilidade. As variáveis $y_1$ e $y_2$ são ditas independentes se e somente se:

$$p(\mathrm{y}_1, \mathrm{y}_2) = p_1(\mathrm{y}_1)p_2(\mathrm{y}_2). \tag{12}$$

Em palavras, a densidade conjunta $p_{y_1y_2}(\mathrm{y}_1, \mathrm{y}_2)$ de $\mathrm{y}_1$ e $\mathrm{y}_2$ deve ser fatorada nos produtos das densidades marginais $p_1(\mathrm{y}_1)$ e $p_2(\mathrm{y}_2)$. Se duas variáveis são independentes, também são descorrelacionadas, mas o contrário não é verdadeiro. Duas variáveis aleatórias serão descorrelacionadas se a covariância $c_{y_1y_2}$ é zero:

$$C_{y_1,y_2} = E\{y_1y_2\} - E\{y_1\}E\{y_2\} = 0 \tag{13}$$

ou equivalentemente,

$$R_{y_1,y_2} = E\{y_1y_2\} = E\{y_1\}E\{y_2\} \tag{14}$$

#### 3.2.3.3. Técnicas de estimação das componentes

O algoritmo fastICA foi inventando por Aapo Hyvärinen (HYVARINEN, 1999) e é eficiente e popular para a análise de componentes independentes. É baseado em um esquema de interação de ponto fixo maximizando a não-gaussianidade como uma medida de independência estatística.

A não-gaussianidade é um elemento chave para a estimação do modelo de ICA, pois a matriz **A** não é identificável quando mais de uma das componentes independentes têm distribuição gaussiana. Considere-se que **x** é distribuído de acordo com o modelo de ICA na equação 11, e que todas as componentes independentes têm distribuições iguais. Para estimar uma das componentes independentes, basta encontrar as combinações lineares de $x_i$, de modo que

$$\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}, \tag{15}$$

Assim, pode-se expressar uma combinação linear de $\mathbf{x_i}$ por:

$$\begin{aligned} y &= \mathbf{b}^{\mathrm{T}}\mathbf{x}, \\ y &= \sum_{\mathrm{i}} b_i x_i, \\ \mathrm{y} &= \mathbf{b}^{\mathrm{T}}\mathbf{A}\mathbf{s}, \end{aligned} \tag{16}$$

em que o vetor **b** deve ser determinado. A partir da equação 16 observa-se que $y$ é uma combinação linear de $s_i$, com coeficientes dados por $\mathbf{q} = \mathbf{b}^{\mathrm{T}}\mathbf{A}$. Logo, tem-se:

$$\begin{aligned} y &= \mathbf{q}^{\mathrm{T}}\mathbf{s}, \\ y &= \sum_{i} q_i s_i. \end{aligned} \tag{17}$$

Se **b** corresponder a uma das linhas da inversa de **A**, então y será uma das componentes independentes e, nesse caso, apenas um dos elementos de $q$ será igual a 1, enquanto todos os outros serão iguais a zero. Não é possível determinar **b** exatamente, contudo, estima-se seu valor com boa aproximação.

Uma forma de determinar $\mathbf{b}$ é variar os coeficientes em $\mathbf{q}$ e então verificar como a distribuição de $y = \mathbf{q}^{\mathbf{T}}\mathbf{s}$ muda. Conforme o Teorema do Limite Central (PAPOULIS, 2002), a soma de variáveis aleatórias aproxima-se cada vez mais de uma distribuição normal então, $y = \mathbf{q}^{\mathbf{T}}\mathbf{s}$ normalmente é mais gaussiana do que qualquer uma das $s_i$ e menos gaussiana quando se iguala a uma das $s_i$. (Note que isto é estritamente verdadeiro apenas se $s_i$ tem distribuições idênticas, como assumimos aqui.) Neste caso, apenas um dos elementos $q_i$ de $\mathbf{q}$ é diferente de zero (HYVÄRINEN; KARHUNEN; OJA, 2001). Como, na prática, os valores de $\mathbf{q}$ são desconhecidos e sabe-se, através das Equações 16 e 17, que:

$$\mathbf{b}^{\mathrm{T}}\mathbf{x} = \mathbf{q}^{\mathrm{T}}\mathbf{s} \tag{18}$$

Pode-se variar $\mathbf{b}$ e observar a distribuição de $\mathbf{b}^{\mathbf{T}}\mathbf{x}$. Portanto, pode-se tomar $\mathbf{b}$ como um vetor que maximiza a não-gaussianidade de $\mathbf{b}^{\mathbf{T}}\mathbf{x}$. Este vetor necessariamente corresponde a $\mathbf{q} = \mathbf{A}^{\mathbf{T}}\mathbf{s}$, o qual possui apenas uma de suas componentes diferente de zero. Isso significa que $y = \mathbf{b}^{\mathbf{T}}\mathbf{x} = \mathbf{q}^{\mathbf{T}}\mathbf{s}$ é igual a uma das componentes independentes. Logo, a maximização da não-gaussianidade de $\mathbf{b}^{\mathbf{T}}\mathbf{x}$ permite encontrar uma das componentes.

#### 3.2.3.4. Negentropia como medida de não-gaussianidade

Uma medida importante de não-gaussianidade é a negentropia. A definição de entropia (HYVÄRINEN; KARHUNEN; OJA, 2001; PAPOULIS, 2002) pode ser generalizada para vetores de variáveis aleatórias contínuas, vindo a ser chamada entropia diferencial. A entropia de uma variável aleatória está relacionada à quantidade de informação que essa variável contém. A entropia será maior quanto mais imprevisível for a variável. Tomando uma variável aleatória $\mathbf{y}$ cuja função densidade de probabilidade é $p_y(\boldsymbol{\eta})$, tem-se a entropia diferencial dada por:

$$H(\mathbf{y}) = -\int p_y(\boldsymbol{\eta}) \log p_y(\boldsymbol{\eta})\, d\boldsymbol{\eta}. \tag{19}$$

Como um dos resultados fundamentais da Teoria da Informação, sabe-se que uma variável gaussiana tem a maior entropia entre todas as variáveis aleatórias de igual variância

(HYVÄRINEN; KARHUNEN; OJA, 2001; PAPOULIS, 2002). Isso quer dizer que uma versão modificada da entropia diferencial pode ser usada como medida de não-gaussianidade. Essa medida é chamada negentropia, definida por:

$$J(\mathbf{y}) = H(\mathbf{y}_{gauss}) - H(\mathbf{y}), \tag{20}$$

em que $\mathbf{y}_{gauss}$ é uma variável aleatória de mesma matriz de correlação (e covariância) que **y**. A negentropia é sempre não negativa e tem valor igual a zero se e somente se **y** tem distribuição gaussiana e é invariante para transformações lineares invertíveis.

Em contraste às suas qualidades como medida de não-gaussianidade, a negentropia é de difícil estimação. Por esta razão, é necessária a utilização de aproximações usando momentos de alta ordem, como por exemplo:

$$J(y) \approx \frac{1}{12} E\{y^3\}^2 + \frac{1}{48} kurt(y^2), \tag{21}$$

onde, $kurt(y)$, a curtose de $y$, é definida como o momento de quarta ordem da variável aleatória $y$, expresso por:

$$kurt(y) = E\{y^4\} - 3(E\{y^2\})^2. \tag{22}$$

A kurtose é zero para variáveis gaussianas e maior que zero para a maioria das variáveis aleatórias não-gaussianas.

### 3.3. Algoritmos de Agrupamento

Os algoritmos de agrupamento têm sido frequentemente utilizados em tarefas de exploração de dados e extração de conhecimento, como detecção de características, segmentação de imagens e outras aplicações em bioinformática. Os resultados obtidos por meio dessa técnica de aprendizado de máquina não supervisionado são altamente dependentes da escolha de parâmetros como as medias de similaridade e os métodos de agrupamentos. Os

algoritmos mais comuns, que também são abordados neste trabalho, é o *k-means*, o nebuloso *c-means* e o mapa auto-organizável, descritos nas seções seguintes.

### *3.3.1. K-means*

A técnica *k-means* (BISHOP, 2006; GAN; MA; WU, 2007) é uma das mais utilizadas, e usa algoritmos de aprendizagem não supervisionados que resolvem o problema do agrupamento. O procedimento segue uma maneira simples e de fácil utilização para classificar um determinado conjunto de dados através de certo número de aglomerados $k$ fixado a priori.

A ideia principal é definir $k$ centroides, um para cada agrupamento. Estes centroides devem ser colocados de uma forma criteriosa, pois localizações diferentes provocam resultados diferentes. Assim, a melhor escolha é colocá-los o mais distante possível um dos outros.

O próximo passo é associar cada ponto pertencente aos dados ao centroide mais próximo. Quando nenhum ponto está pendente, a primeira etapa está concluída e um agrupamento preliminar foi realizado. Neste ponto, precisa-se voltar a calcular $k$ novos centroides como baricentros dos agrupamentos resultantes da etapa anterior.

Em seguida uma nova ligação tem de ser feita entre os mesmos pontos de conjunto de dados e os novos centroides mais próximos. Repetindo esse processo várias vezes, têm-se como resultado que os $k$ centroides mudam a sua localização passo a passo até que não haja mais mudanças. Em outras palavras, os centroides não se movem mais. Finalmente, este algoritmo visa minimizar uma função objetivo, neste caso uma função de erro quadrado. A função objetivo é:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - c_k\|^2 \qquad (23)$$

onde $N$ é a quantidade de pontos pertencentes a $x$, $K$ é a quantidade de agrupamentos definidos pelo usuário, $r_{nk} \in \{0,1\}$ corresponde a um indicador binário, para determinar se o ponto $x_n$ corresponde ou não ao centróide $c_k$, correspondente ao agrupamento $k$.

### *3.3.2.* Nebuloso *c-means*

Em algoritmos de agrupamento convencionais cada elemento pertence somente a um grupo e todos os grupos são considerados diferentes entre si. Em alguns casos, os conjuntos não são completamente diferentes e um elemento pode ser considerado pertencente a um grupo tanto quanto a outro. Esta situação não pode ser caracterizada com processos de classificação clássicos. Então a diferença entre os conjuntos tem uma noção nebulosa e os dados que representam este tipo de estrutura são mais precisamente manipulados por métodos de agrupamento nebuloso.

O algoritmo nebuloso *c-means* (FCM, do inglês *fuzzy c-means*) é o mais popular, amplamente utilizado e eficiente método de agrupamento que utiliza o agrupamento nebuloso, onde cada elemento pode pertencer a todos os grupos, com diferentes graus de pertinência entre zero a um. Este algoritmo é baseado na minimização da seguinte função objetivo (HATHAWAY; BEZDEK, 2008; DUNN, 1973):

$$J_m(U, v) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m \|x_k - v_i\|^2, \tag{24}$$

onde, $x_k$ são os vetores de amostras, $v_i$ são os centros dos grupos, $U = \{u_{ik}\}$ é uma matriz $c \times n$ onde $u_{ik}$ é o $i$-ésimo valor de pertinência da $k$-ésima amostra de entrada $x_k$, e os valores de pertinência devem satisfazer as seguintes condições:

$$\begin{gathered} 0 < u_{ik} < 1 \; \forall \; 1 \leq i \leq c \; e \; 1 \leq k \leq n, \\ \sum_{i=1}^{c} u_{ik} = 1 \;\; \forall \; 1 \leq k \leq n, \\ \sum_{k=1}^{n} u_{ik} > 0 \; \forall \; 1 \leq i \leq c \quad e \quad 1 \leq m < \infty. \end{gathered} \tag{25}$$

A função objetivo é a soma do quadrado das distâncias euclidianas de cada elemento e centro do agrupamento. As distâncias são ponderadas pelos graus de pertinências. O algoritmo é interativo e utiliza as seguintes equações:

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m},$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left[ \frac{\|x_k - v_i\|^2}{\|x_k - v_j\|^2} \right]^{1/(m-1)}}. \quad (26)$$

Para o cálculo do centro de um agrupamento, todos os dados de entrada são considerados e as suas contribuições são ponderadas pelos valores de pertinência. Para cada dado, seu valor de pertinência em cada agrupamento depende da distância ao centro do agrupamento. O fator de peso $m$ reduz a influência de valores de pertinência pequenos. Para grandes valores de $m$, menores são as influências de dados com pequenos valores de pertinência.

### 3.3.3. Mapa auto-organizável

O mapa auto-organizável (SOM, do inglês *self-organizing maps*) é uma rede neural não supervisionada que realiza agrupamentos por meio de aprendizagem competitiva (KOHONEN, 1990; HAYKIN, 2009). No SOM os neurônios são normalmente dispostos numa estrutura bidimensional (o mapa de características) e cada neurônio recebe a informação a partir da camada de entrada e a partir dos outros neurônios no mapa.

O algoritmo responsável pela formação do mapa auto-organizável inicia atribuindo valores aleatórios aos pesos sinápticos $\mathbf{w}_j$ da rede. Após a inicialização da rede, três processos essenciais estão envolvidos na formação do SOM: competição, cooperação e adaptação sináptica.

No processo de competição para cada padrão de entrada os neurônios na rede calculam os respectivos valores de uma função discriminante, equação 27. Esta função discriminante fornece a base para a competição entre os neurônios. O neurônio com o maior valor de função discriminante é declarado vencedor da competição.

$$i(\mathbf{x}) = arg\ \min_j \|\mathbf{x} - \mathbf{w}_j\|. \tag{27}$$

O processo seguinte é o da cooperação, onde o neurônio vencedor determina a localização espacial de uma vizinhança topológica, $h_{j,i}$, de neurônios excitados, fornecendo assim a base para a cooperação entre esses neurônios vizinhos:

$$h_{j,i(x)} = exp\left(-\frac{d_{j,i}^2}{2\sigma^2(n)}\right), \tag{28}$$

onde $d_{j,i}$ é a distância lateral entre o neurônio vencedor $i$ e o neurônio excitado $j$ e $\sigma(n)$ é dado pela equação 29:

$$\sigma(n) = \sigma_0 exp\left(\frac{n}{\tau_1}\right), \tag{29}$$

onde $\sigma_0$ é o valor de $\sigma$ na inicialização do algoritmo SOM e $\tau_1$ é uma constante de tempo a ser escolhida pelo usuário.

A última etapa na formação do mapa auto-organizável é o processo adaptativo, onde o vetor de peso sináptico $\mathbf{w}_j$ do neurônio $j$ no tempo $n$ é obrigado a mudar em relação ao vetor $\boldsymbol{x}$ através da equação 30:

$$\mathbf{w_j}(n+1) = \mathbf{w_j}(n) + \eta(n) h_{j,i(x)}(n)\left(x(n) - \mathbf{w_j}(n)\right), \tag{30}$$

onde

$$\eta(n) = \eta_0 exp\left(-\frac{n}{\tau_2}\right), \tag{31}$$

e $\tau_2$ é outra constante de tempo.

### 3.4. Análise discriminante linear

A análise discriminante linear (LDA, do inglês, *linear discriminant analysis*), como o nome sugere, busca por uma combinação linear de variáveis de entrada que podem proporcionar uma separação adequada para as classes dadas. Ao invés de olhar para uma determinada forma de distribuição paramétrica, a LDA utiliza uma aproximação empírica para definir um plano de decisão linear no espaço dos atributos, i.e., modelos de superfícies. A função discriminante usada pelo LDA é construída como uma combinação linear das variáveis que tentam de alguma forma maximizar as diferenças entre as classes (LACHENBRUCH, 1975; BISHOP, 2006), conforme:

$$\mathbf{y} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}. \tag{32}$$

Existem inúmeras variações para se encontrar o vetor $\boldsymbol{\beta}$, uma das mais bem sucedidas é a regra do discriminante linear de Fisher. A regra de Fisher é considerada uma classificação "sensível" no sentido de que é intuitivamente atraente. Ela faz uso do fato de que as distribuições com uma maior variância entre as classes, do que dentro de cada classe, é mais fácil de separar. Por isso, ela procura por uma função linear no espaço de atributos que maximiza a razão da soma dos quadrados entre-grupos ($B$) e a da soma dos quadrados intra-grupos ($W$). Isso pode ser conseguido através da maximização de:

$$\mathbf{S} = \frac{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{B}\boldsymbol{\beta}}{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{W}\boldsymbol{\beta}}, \tag{33}$$

O vetor $\boldsymbol{\beta}$ que maximiza esta razão é o autovetor correspondente ao maior autovalor de $\mathbf{W}^{-1}\mathbf{B}$, isto é, a função discriminante linear $\mathbf{y}$ é equivalente a primeira variável canônica. A regra do discriminante pode ser escrita como:

$$\mathbf{x} \in i \; se \; |\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{u}_i| < \left|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{u}_j\right| \; \forall \, j \neq i, \tag{34}$$

onde, $\mathbf{W} = \sum \mathbf{P}_i \boldsymbol{\Sigma}_i$ e $\mathbf{B} = \sum \mathbf{P}_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$ e $\mathbf{P}_i$ é o tamanho da amostra da classe $i$, $\boldsymbol{\Sigma}_i$ é a matriz de covariância da classe $i$, $\mu_i$ é valor médio da amostra da classe $i$ e $\mu$ é a média populacional.

# 4. MÉTODO E RESULTADOS DA SEGMENTAÇÃO

Neste capítulo é proposta uma metodologia para a detecção de regiões de interesse em mamografias digitalizadas, a partir de imagens bases geradas pelas técnicas de extração de características: *wavelets* de Gabor, análise de componentes principais e análise de componentes independentes. Para o agrupamento utiliza-se dos algoritmos de *k-means*, nebuloso *c-means* e mapas auto-organizáveis. Nestas ROIs pode haver nódulos benignos ou malignos, então esta metodologia serviria como um pré-processamento para a classificação apresentada posteriormente.

## 4.1. Introdução

O diagrama de blocos do método proposto é ilustrado na Figura 15. Consiste basicamente em filtrar a imagem original, usando um banco de filtros extraídos das técnicas de *wavelets* de Gabor, PCA e ICA. Essas imagens filtradas são associadas pelos algoritmos de agrupamento *k-means*, nebuloso *c-means* e mapas auto-organizáveis, resultando em uma imagem segmentada.

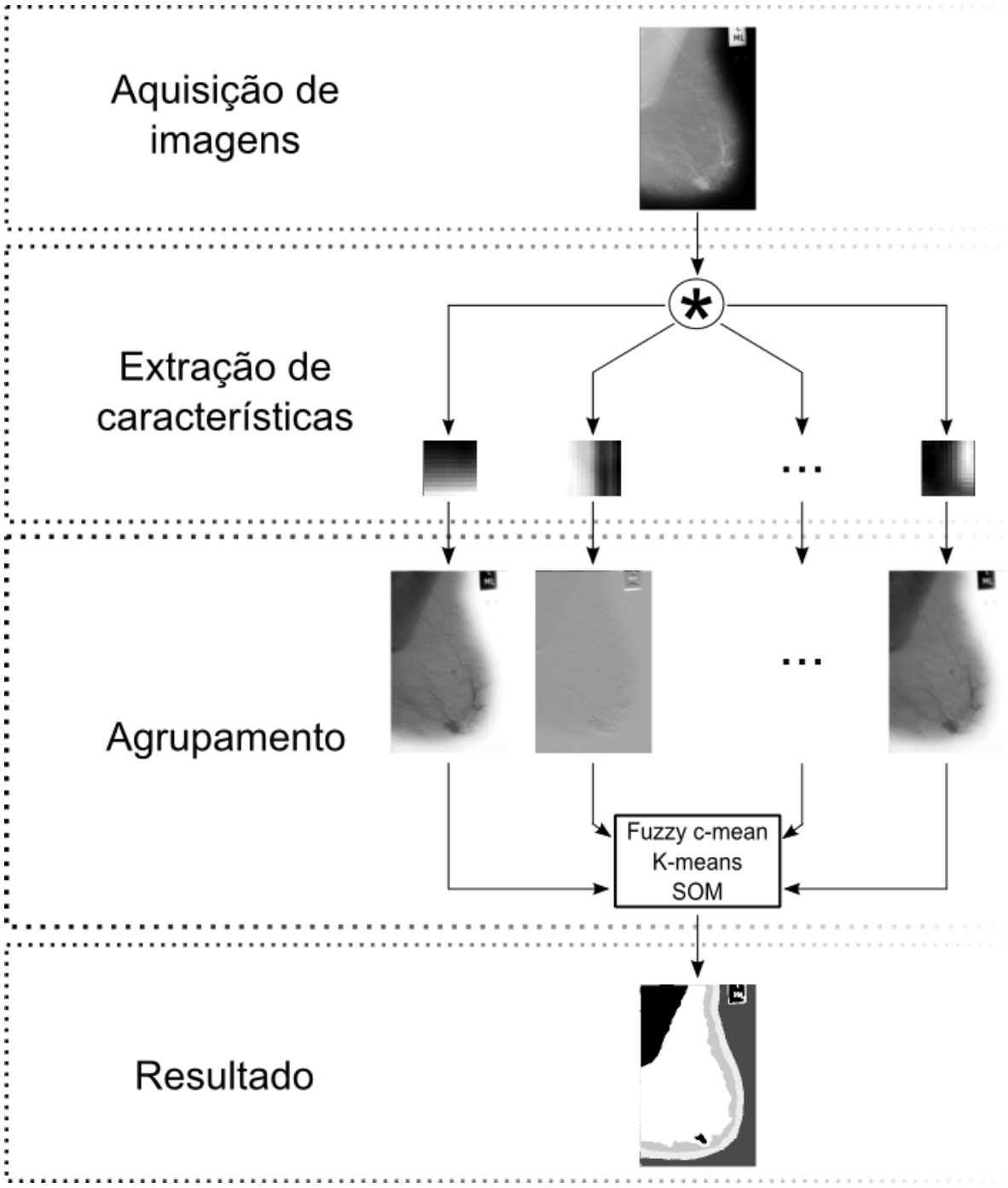


**Figura 15: Metodologia proposta em três passos para segmentação de massas em mamografias digitalizadas: aquisição de imagens; extração de características; agrupamento por *k-means*, nebuloso *c-means* e mapa auto-organizável.**

### 4.2. Aquisição de Imagens

Para o desenvolvimento e avaliação da metodologia proposta, foi usada uma base de dados disponível publicamente: *mini-MIAS* (*Mammographic Image Analysis Society*) *database* (SUCKLING et. al., 1994).

A base de dados mini-MIAS foi fornecida pela *Mammografic Institute Analysis Society* (MIAS) (SUCKLING et. al., 1994). As mamografias têm um tamanho de 1024x1024 pixels e resolução de 200 micron. Esta base de dados é composta por 332 mamogramas, tanto de mamas do lado direito quanto do lado esquerdo. Destas pacientes, 51 são diagnosticadas como malignos, 67 como benignos e 211 normais. Cada uma das anormalidades é classificada conforme seu tipo: calcificação, massas circunscritas, distorções assimétricas da arquitetura e outros. Para reduzir os ruídos, cada imagem foi submetida a um filtro de média de tamanho $3 \times 3$, conforme é ilustrado na Figura 16(a), imagem original, e Figura 16(b), imagem filtrada.

Em seguida, convertemos cada mamografia de teste em uma imagem binária, Figura 16(c), utilizando-se o limiar global obtido pelo método de Otsu (OTSU, 1979). Este método baseia-se na análise discriminante. A operação de limiarização é considerada como sendo o particionamento dos pixels de uma imagem em duas classes $C_0$ e $C_1$, que representam o objeto e o fundo, ou vice-versa, sendo que esta partição dar-se-á no nível de cinza $t$.

Um limiar ótimo pode ser obtido através da maximização da seguinte função:

$$\rho = \frac{{\sigma_B}^2}{{\sigma_T}^2}. \tag{35}$$

Onde,

$${\sigma_B}^2 = \omega_0 \omega_1 (\mu_1 - \mu_0)^2 \tag{36}$$

$${\sigma_T}^2 = \sum_{i=1}^{L} (i - \mu_T)^2 p_i \tag{37}$$

$$\omega_0 = Pr(C_0) = \sum_{i=1}^{t} p_i = \omega(t) \tag{38}$$

$$\omega_1 = Pr(C_1) = \sum_{i=t+1}^{L} p_i = 1 - \omega(t) \tag{39}$$

$$\mu_0 = \sum_{i=1}^{t} i\ Pr(i|C_0) = \sum_{i=1}^{t} \frac{i\ p_i}{\omega_0} = \frac{\mu(t)}{\omega(t)} \tag{40}$$

$$\mu_1 = \sum_{i=t+1}^{L} i\ Pr(i|C_1) = \sum_{i=t+1}^{L} \frac{i\ p_i}{\omega_1} = \frac{\mu_T - \mu(t)}{1 - \omega(t)} \tag{41}$$

$$\mu_T = \mu(L) = \sum_{i=1}^{L} i\ p_i \tag{42}$$

Utilizou-se o operador de abertura para suavizar contornos, quebrar istmos estreitos e eliminar pequenas ilhas e picos agudos das imagens binarizadas, Figura 16(d), auxiliando para um melhor resultado no agrupamento das imagens e remoção de artefatos.

A obtenção do operador de abertura é realizada por dois processos na seguinte ordem: Dilatação, dada pelo operador $\oplus$, responsável pelas expansões das fronteiras da imagem e Erosão, dada pelo operador $\ominus$, responsável pela contração das fronteiras da imagem. Desta forma o operador de abertura pode ser dado pela seguinte equação:

$$\boldsymbol{I} \circ \boldsymbol{S} = (\boldsymbol{I} \oplus \boldsymbol{S}) \ominus \boldsymbol{S} \tag{43}$$

onde, $\boldsymbol{I}$ é a imagem original e $\boldsymbol{S}$ é uma estrutura em forma de disco com o raio de 5 pixels, dada pela matriz:

$$\boldsymbol{S} = \begin{matrix} 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \end{matrix} \quad (44)$$

Observa-se então que além da região de mama outras estruturas foram detectadas. Para a remoção destes artefatos busca-se a estrutura de maior área localizada na imagem e excluem-se todas as outras, Figura 16(e). Por último recuperamos os valores de cada pixel na posição em que a mama foi localizada, Figura 16(f).

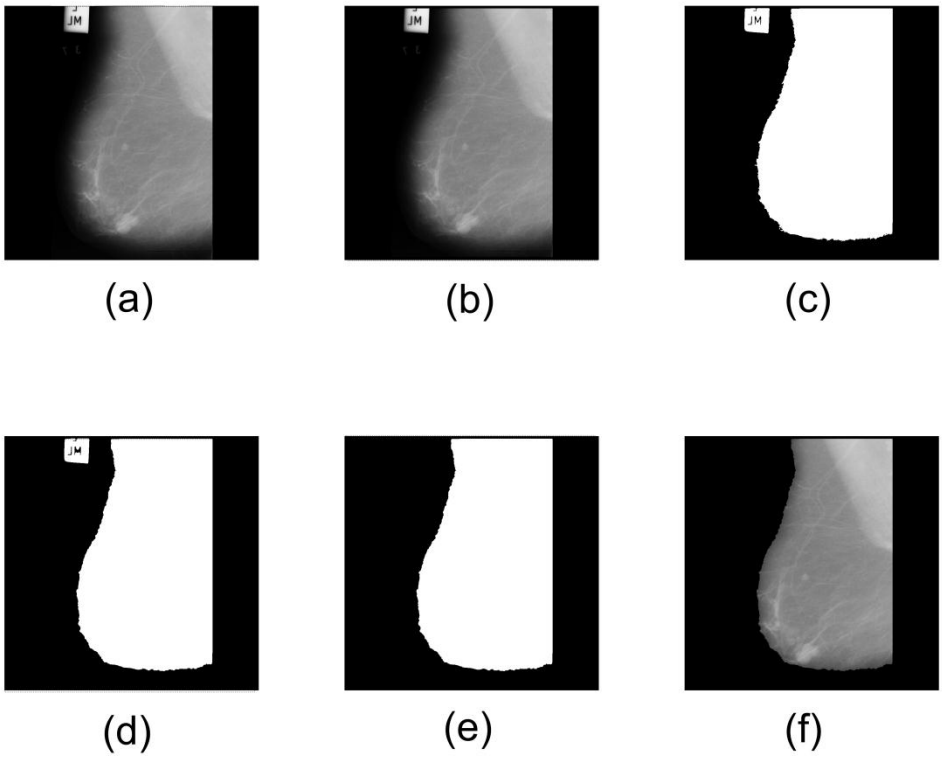


**Figura 16: Passo a passo da remoção de ruídos e artefatos das mamografias. Em (a) tem-se a imagem mdb005 do banco de dados MIAS (SUCKLING et. al., 1994). Em (b) tem-se a imagem original filtrada pelo filtro de média. (c) É a imagem binarizada utilizando o limiar global do método de Otsu. (d) É a imagem binária após a operação de abertura. (e) Ilustra o passo da remoção de artefatos da mamografia. (f) Resultado deste processo de remoção de ruídos e artefatos.**

### 4.3. Extração de Características

O segundo passo desta metodologia é a extração de características. São utilizadas oito imagens bases de três diferentes métodos de extração de características: *wavelets* de Gabor, análise de componentes principais (PCA) e análise de componentes independentes (ICA). Esta quantidade de componentes foi escolhida através de um experimento empírico (CAMPOS; COSTA; BARROS, 2008) em que foi observado que o melhor resultado utiliza-se de oito imagens bases com um tamanho de $20 \times 20$ *pixels*.

Pelas equações 1 e 2 (seção 2.3.1), foram geradas oito *wavelets* de Gabor, com quatro diferentes orientações (0°, 45°, 90° e 135º) e duas frequências (1,25 Hz e 10 Hz) a uma taxa de amostragem de 20 Hz.

Para PCA e ICA, foram retirados *patchs* de tamanho $20 \times 20$ de 88 regiões de massa para o treinamento, ou seja, das imagens que não fizeram parte do teste. Foram selecionadas as oito primeiras componentes principais, i.e., as oito componentes de maior variância. Na Tabela 3 podem ser visto os valores das variâncias de cada componente assim como as variâncias acumuladas. Neste experimento as oito componentes conseguem explicar um total de 96% da variância dos dados. As outras 30 ROIs foram utilizadas como teste, para avaliar o sistema de segmentação.

**Tabela 3: Variância de cada componente principal e a variância acumulada, na redução de dimensionalidade por PCA.**

| Componente | Variância (%) | Variância acumulada (%) |
|---|---|---|
| 1 | 76,39 | 76,39 |
| 2 | 7,34 | 83,73 |
| 3 | 5,68 | 89,41 |
| 4 | 2,22 | 91,63 |
| 5 | 2,06 | 93,69 |
| 6 | 0,92 | 94,61 |
| 7 | 0,86 | 95,46 |
| 8 | 0,59 | 96,06 |

Para encontrar as funções bases, foi utilizado o algoritmo fastICA, onde o critério de parada precisava atender dois requisitos: $|W_{atual} - W_{anterior}| < 10^{-50}$, i.e., invariância das direções dos vetores colunas da matriz $W$; e um mínimo de 1000 interações. Foi necessária uma redução de dimensionalidade por PCA nos dados para se obter apenas as oito funções bases.

As Figuras 16, 17 e 18 representam respectivamente os filtros obtidos por Gabor, PCA e ICA.

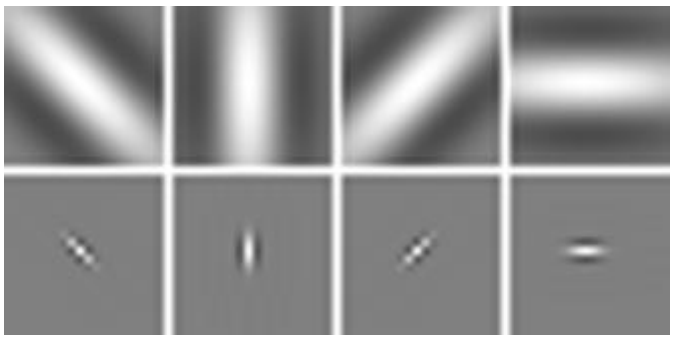

**Figura 17: Oito filtros de Gabor utilizados para a filtragem das imagens.**

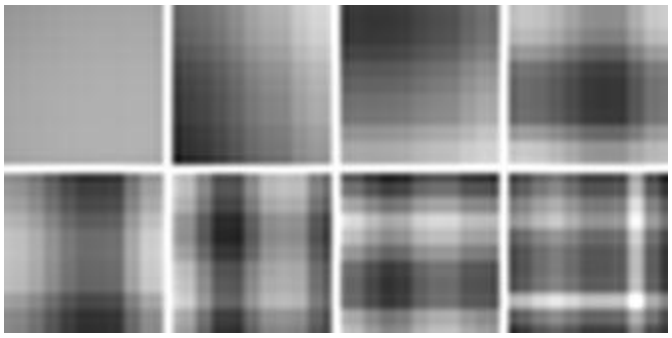

**Figura 18: Oito filtros da PCA utilizados para a filtragem das imagens.**

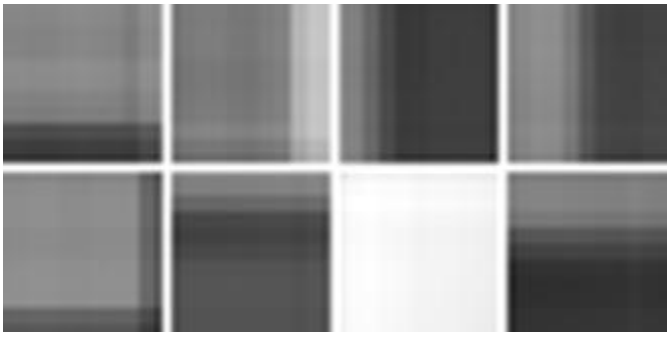

**Figura 19: Oito filtros da ICA utilizados para a filtragem das imagens.**

Após a obtenção de todas as imagens bases, foi realizada uma filtragem das mamografias de teste com as imagens bases obtidas pelos métodos de extração de características. Matematicamente, uma filtragem é uma convolução, dada pela equação 45.

$$\mathbf{Y}(n_1, n_2) = \mathbf{I}(n_1, n_2) * \mathbf{H}(n_1, n_2) = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \mathbf{I}(k_1, k_2)\mathbf{H}(n_1 - k_1, n_2 - k_2) \quad (45)$$

onde, **I** é a imagem original; **H** é o filtro e **Y** é a imagem processada. A Figura 20 ilustra a convolução de uma mamografia com os oito filtros gerados pela ICA.

### 4.4. Agrupamento

Após o processo de filtragem, foram obtidas oito imagens filtradas, uma para cada filtro obtido pela extração de características. Para uma análise coerente dos dados obtidos torna-se necessário realizar uma associação das imagens filtradas em uma única imagem, com o intuito de identificar e isolar as imagens em regiões de interesse que possam corresponder a algum tipo de massa. Este processo é denominado agrupamento.

O agrupamento é feito através de algoritmos como *k-means*, nebuloso *c-means* e o *SOM*. Estes algoritmos recebem várias imagens e associam características comuns em uma única imagem, rotulando-os com o seu respectivo conjunto, produzindo finalmente a imagem segmentada. O número de classes que ser quer agrupar deve ser pré-definido e, de forma empírica, os melhores resultados foram obtidos utilizando cinco classes. Todos os algoritmos utilizados para o agrupamento neste trabalho foram relatados na seção 3.3 – Algoritmos de Agrupamento.

Para uma redução do número de falsos positivos foram desconsideradas regiões muito pequenas, áreas que contenham menos de 50 *pixels*, e estruturas que não tivessem nenhuma proximidade com uma circunferência, ou seja, tivessem uma circularidade maior que 5. Como medida de circularidade utilizou-se a equação 46.

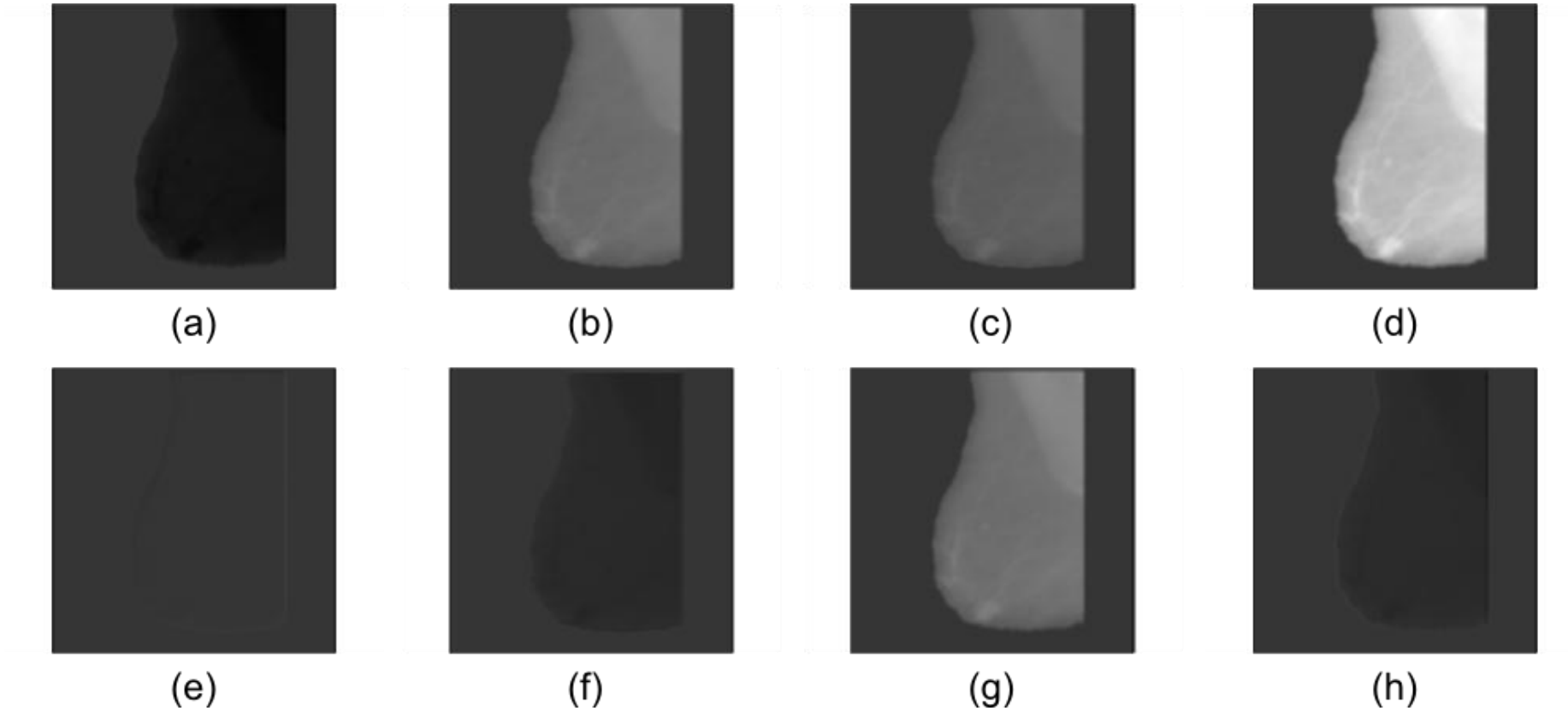


**Figura 20: Convolução da imagem com os filtros gerados pela extração de características. Nesta figura a imagem original foi convoluída para cada um dos oito filtros da ICA, resultando assim em oito imagens que devem ser associadas pelos algoritmos de agrupamento.**

$$circularidade = \frac{(perímetro)^2}{4\pi \times área}, \tag{46}$$

nesta equação um círculo perfeito teria uma $circularidade$ igual a 1.

Em seguida são isoladas as regiões da mesma classe do músculo do peito, já que as intensidades dos *pixels* desta área são próximas dos valores de *pixel* dos nódulos. Por último é retirado o músculo peitoral maior da imagem, ficando apenas a ROI, conforme é ilustrada na Figura 21.


**Figura 21: Localização das regiões de interesse. Em (a) tem-se a imagem original. (b) Ilustra o resultado do agrupamento realizado pelo *k-means*, isolando-se apenas o grupo da região do músculo. (c) É a região de interesse que foi obtida pela retirada da região do músculo peitoral maior.**

### 4.5. Avaliação do Método

Para avaliar a segmentação em relação à sua capacidade de detectar regiões de massa, foram analisados a sua fração de localização de lesão (FLL) e a fração de não lesão localizada (FNL).

A FLL (onde, $0 \leq FLL \leq 1$) indica o quanto é bom o método para identificar regiões de massa e é definida por:

$$FLL = \frac{LL}{NL}, \tag{47}$$

onde LL é o número de lesões localizadas e NL é o número de lesões, ou seja, a quantidade total de regiões de massa.

A FNL (onde, $0 \leq FNL$) indica a fração de quantas regiões de não-massas foram detectadas como possíveis regiões de massa. Quanto menor o FNL melhor será o método. A FNL é definida por:

$$FNL = \frac{NNL}{NI}, \tag{48}$$

onde NNL é o número de não-lesões, isto é, a quantidade total de regiões de massa e NI é o número total de imagens que foram segmentadas.

### 4.6. Resultados

Os resultados utilizando os algoritmos *k-means*, nebuloso *c-means* e mapa auto-organizável estão elencados nas tabelas 4 e 5.

**Tabela 4: Quantidades de verdadeiros positivos (VP) e falsos positivos (FP) de um total de 30 imagens para cada técnica e seu respectivo algoritmo de agrupamento.**

| | ***K-means*** | | | **Nebuloso *c-means*** | | | **Mapa auto-organizável** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Gabor** | **PCA** | **ICA** | **Gabor** | **PCA** | **ICA** | **Gabor** | **PCA** | **ICA** |
| **VP** | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 | 29 |
| **FP** | 120 | 67 | 64 | 121 | 79 | 71 | 142 | 80 | 79 |

**Tabela 5: Avaliação da metodologia da segmentação através de suas medidas: fração de localização de lesão (FLL) e fração de não lesão localizada (FNL).**

| | ***K-means*** | | | **Nebuloso *c-means*** | | | **Mapa auto-organizável** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Gabor** | **PCA** | **ICA** | **Gabor** | **PCA** | **ICA** | **Gabor** | **PCA** | **ICA** |
| **FLL** | 0,96 | 0,96 | 0,96 | 0,96 | 0,96 | 0,96 | 0,96 | 0,96 | 0,96 |
| **FNL** | 4,00 | 2,23 | 2,13 | 4,03 | 2,63 | 2,36 | 4,73 | 2,66 | 2,63 |

Para a detecção de verdadeiros positivos, todas as técnicas utilizadas neste trabalho cometeram apenas um erro nas 30 imagens analisadas. Na quantidade de falsos positivos, o *k-means* obteve o melhor resultado, chegando a acusar de 64 a 120 falsos positivos dependendo da técnica de extração de características. O nebuloso *c-means* encontrou de 71 a 121 e o mapa auto-organizável foi o que encontrou a maior quantidade de falsos positivos, variando entre 79 a 142.

Em relação à validação da metodologia através da FLL e da FNL, pode-se observar que a FLL atingiu 0,96 para todos os algoritmos de agrupamento, independente da técnica de extração de característica. A FNL do *k-means* foi a que obteve o melhor resultado, variando entre 2,13 a 4,00, dependendo da extração de características. O nebuloso *c-means* ficou em segundo atingindo uma FNL entre 2,36 a 4,03 e o mapa auto-organizável obteve o pior resultado, alcançando uma FNL de 2,63 a 4,73.

Validando estes resultados através das técnicas de extração de características, observa-se que apesar de todas encontrarem as mesmas quantidades de lesões, o número de FPs variou muito entre uma técnica e outra. As *wavelets* de Gabor foram as que obtiveram o pior resultado, variando de 120 a 142, dependendo do algoritmo de agrupamento. A PCA ficou em segundo lugar, variando de 67 a 80 e a ICA foi a que obteve o melhor resultado, variando entre 64 a 79.

Desta forma, observa-se que o melhor resultado obtido foi com o algoritmo de agrupamento *k-means* e com a técnica de análise de componentes independentes como extrator de característica, atingindo uma FLL de 0,96 e uma FNL de 2,13.

Na Figura 22 é ilustrado um dos resultados obtidos por esta metodologia. Podem-se comparar os três métodos de codificação juntamente com os três algoritmos de agrupamento descrito neste trabalho. Observa-se que o resultado é a localização precisa da região de massa. Neste caso não foi gerado nenhum falso positivo para os conjuntos de técnicas de PCA e ICA, já as *wavelets* de Gabor geraram dois falsos positivos para cada algoritmo de agrupamento.

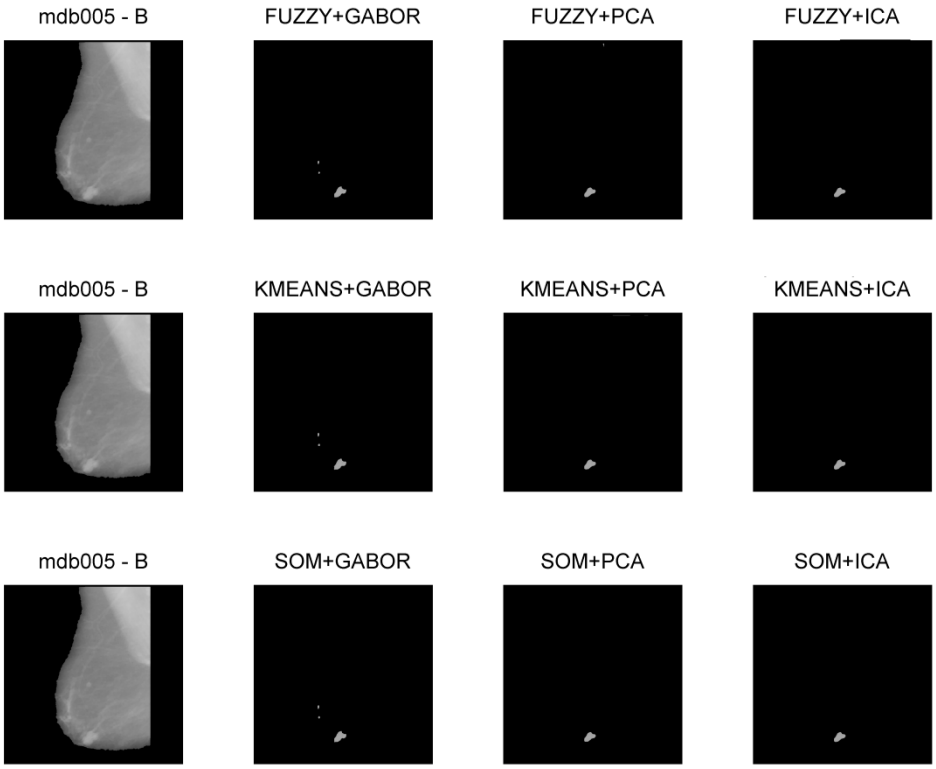


**Figura 22: Imagem MDB005 do MIAS segmentada pelos três algoritmos de agrupamento e as três técnicas de codificação.**

Na Figura 23, além da identificação do nódulo, pode-se observar que foram detectadas pequenas regiões de massa em algumas técnicas, incrementando a taxa de falsos positivos. Através de uma análise visual um radiologista poderia facilmente descartar esses FP e analisar apenas a maior região de massa, que é um nódulo benigno. Em termos computacionais, um classificador como a LDA descartaria estas regiões de FP, pois elas não contêm as características necessárias para serem classificadas como uma região de massa. Nesta figura, observa-se que as *wavelets* de Gabor foram as que encontraram a maior quantidade de falsos positivos, três, a PCA encontrou duas regiões de falsos positivos e a ICA não encontrou nenhum FP.

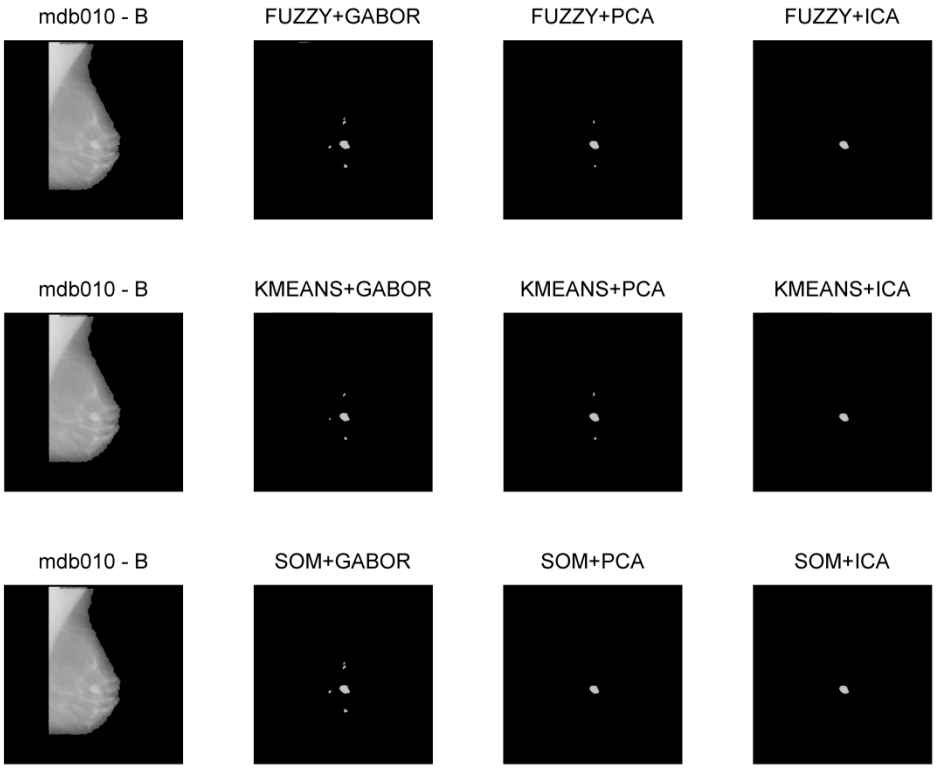


**Figura 23: Imagem MDB010 segmentada pelos três algoritmos de agrupamento e as três técnicas de codificação.**

Na Figura 24, pode-se observar que a metodologia tem uma dificuldade em detectar apenas os nódulos em mamas densas, aumentando consideravelmente a quantidade de FP, pois estes tecidos têm características visuais muito próximas dos tecidos de nódulos. Este tipo de problema pode ocorrer principalmente em mulheres jovens que tenham as mamas densas, em compensação estas mulheres estão fora do grupo de risco, considerando que a idade é um dos maiores fatores para o desenvolvimento do câncer de mama.

Nesta figura, independente do algoritmo de agrupamentos, observa-se que Gabor obteve o maior número de FP, 13, a PCA obteve 11 e a ICA obteve o menor número de FP, 10.

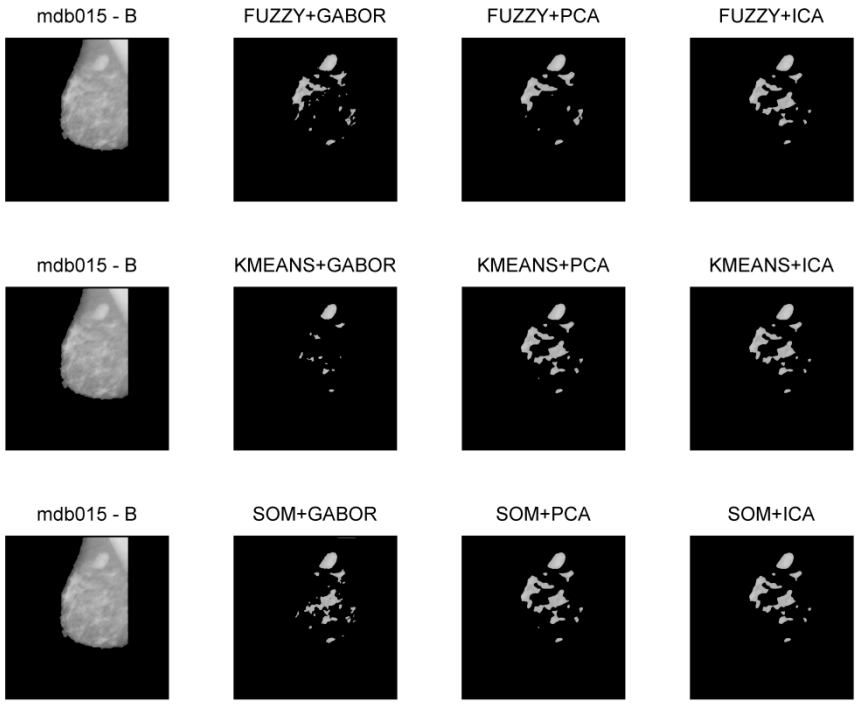


**Figura 24: Imagem MDB017 segmentada pelos três algoritmos de agrupamento e as três técnicas de codificação.**

Outro problema que ocorre pode ser visto na Figura 25, mamas que contenham algum tipo de calcificação também podem apontar como um falso positivo e devem ser analisadas com mais cuidado por um especialista.

Devido a grande dificuldade da detecção de nódulos em mamas densas, foi realizado este mesmo estudo utilizando as 35 mamografias densas do banco de dados MIAS para o teste e todas as outras 83 imagens que continham nódulos foram utilizadas para o treinamento.

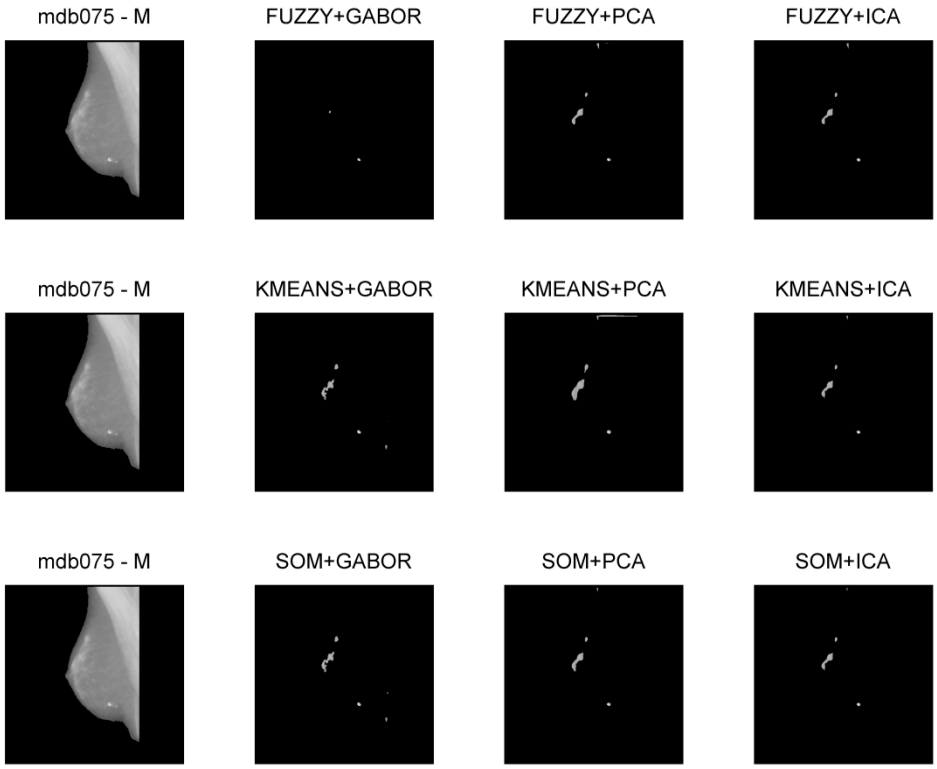


**Figura 25: Imagem MDB075 segmentada pelos três algoritmos de agrupamento e as três técnicas de codificação.**

Apesar de que a mama densa seja uma característica de mulheres jovens, fora do grupo de risco, isso não a impede de desenvolver algum tipo de nódulo, principalmente aquelas que herdaram mutações genéticas nos genes BRCA1 e BRCA2, pois tem o risco muito aumentado de desenvolver câncer de mama. Estes resultados estão apresentados nas tabelas 6 e 7.

**Tabela 6: Quantidades de verdadeiros positivos (VP) e falsos positivos (FP) de um total de 35 imagens de mamas densas para cada técnica e seu respectivo algoritmo de agrupamento.**

| | ***K-means*** | | | **Nebuloso *c-means*** | | | **Mapa auto-organizável** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Gabor** | **PCA** | **ICA** | **Gabor** | **PCA** | **ICA** | **Gabor** | **PCA** | **ICA** |
| **VP** | 23 | 26 | 29 | 24 | 27 | 29 | 24 | 28 | 29 |
| **FP** | 175 | 70 | 98 | 176 | 95 | 103 | 175 | 105 | 103 |

**Tabela 7: Avaliação da metodologia da segmentação através de suas medidas fração de localização de lesão (FLL) e fração de não lesão localizada (FNL) de um total de 35 imagens de mamas densas para cada técnica e seu respectivo algoritmo de agrupamento.**

| | ***K-means*** | | | **Nebuloso *c-means*** | | | **Mapa auto-organizável** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Gabor** | **PCA** | **ICA** | **Gabor** | **PCA** | **ICA** | **Gabor** | **PCA** | **ICA** |
| **FLL** | 0,65 | 0,74 | 0,82 | 0,68 | 0,77 | 0,82 | 0,68 | 0,80 | 0,82 |
| **FNL** | 5,00 | 2,00 | 2,80 | 5,02 | 2,71 | 2,94 | 5,00 | 3,00 | 2,94 |

Observa-se que a ICA obteve os melhores resultados, independente da técnica de agrupamento, sendo que a FLL foi de 0,82 para todos os algoritmos. A PCA obteve um resultado inferior ao da ICA, atingindo uma FLL entre 0,74 a 0,80 e as *wavelets* de Gabor foram as que obtiveram o pior resultado, atingido uma FLL entre 0,65 a 0,68.

Em termos de falsos positivos, Gabor também obteve o pior resultado com uma FNL de aproximadamente 5,0 para todos os algoritmos de agrupamento. A PCA obteve um resultado superior ao de Gabor, atingindo uma FNL entre 2,00 a 3,00, dependendo do algoritmo de agrupamento. A ICA, que obteve uma FNL entre 2,80 a 2,94, não conseguiu gerar uma menor quantidade de falsos positivos em relação ao PCA. Em compensação, a FLL foi 8% superior, o que faz com que a técnica de ICA tenha o resultado mais expressivo do que as outras técnicas.

### 4.7. Discussões e Conclusões

Neste capítulo foi apresentado um sistema de detecção de nódulos em mamografias digitalizadas auxiliado por computador. Foram utilizadas técnicas de codificação como extrator de características (*wavelets* de Gabor, PCA e ICA) em conjunto com os algoritmos de agrupamentos *k-means*, nebuloso *c-means* e mapas auto-organizáveis.

O fator principal para uma segmentação precisa é a extração de características. Pode-se observar nos resultados que as melhores taxas de sucesso ocorrem com a técnica de ICA, independente do algoritmo de agrupamento. É provável que isto ocorra devido as melhores representações da codificação eficiente, garantindo uma independência estatística enquanto a técnica de PCA garante apenas a descorrelação dos dados. A *wavelet* de Gabor foi a técnica que menos encontrou regiões de massa, comparada com as outras técnicas de codificação.

Apesar da diferença no resultado entre uma técnica e outra de codificação, principalmente entre PCA e ICA, ter sido pequena em pontos percentuais na taxa de sucesso, estes percentuais são muito valiosos, pois a ocorrência de falsos positivos (baixa especificidade) pode levar a paciente a uma biópsia desnecessária. No primeiro teste, a diferença da FNL entre PCA e ICA alcançou o valor de 0,1 com *k-means*, 0,33 com nebuloso

*c-means* e 0,03 com SOM. Para o segundo teste, o teste de mamas densas, foi encontrado uma diferença de 0,80 para o *k-means*, 2,31 para o nebulos *c-means* e 0,06 para o SOM.

Em números reais, no primeiro teste, a ICA com *k-means,* gerou 3 falsos positivos a menos que a PCA com este mesmo algoritmo. Com o nebuloso c-means, a PCA gerou 10 FPs a mais que o ICA. Com o SOM, a PCA gerou 1 falso positivo a mais. No segundo teste, a diferença conseguiu ser maior ainda, de 0,80 no *k-means*, representando 28 falsos positivos a mais na PCA que na ICA. Com nebuloso *c-means* a diferença foi de 0,23, 8 FPs, enquanto que no SOM foi de 0,06, representando 2 FPs a mais.

Dos algoritmos de agrupamento, o *k-means* foi o que menos gerou falsos positivos e um dos que mais encontrou verdadeiros positivos, enquanto o mapa auto-organizável foi o que obteve o pior resultado, gerando uma grande quantidade de falsos positivos. O algoritmo de nebuloso *c-means* foi o que obteve um taxa de sucesso intermediária entre os outros dois.

Uma vantagem encontrada neste sistema foi a insensibilidade em relação ao tamanho dos nódulos, já que foram detectadas massas de diferentes tamanhos, incluindo nódulos de pequenas dimensões. Algumas microcalcifações também foram encontradas e, apesar de não serem consideradas massas, essas regiões devem requerer uma atenção do especialista.

Durante o período da realização dos testes, cogitou-se a utilização do algoritmo *Gauss Mean Shift*, um algoritmo de agrupamento em que não é necessário definir o número de grupos, ao contrário dos algoritmos utilizados neste trabalho. Mas esta metodologia tornou-se inviável devido ao longo período de processamento (36 horas em uma única imagem) sem um resultado satisfatório.

Outro teste sem sucesso realizado durante este trabalho está relacionado aos péssimos resultados quando ocorre a remoção do músculo do peito antes do processo de agrupamento. Imaginava-se que ao retirar este tecido, que não contêm regiões de massa, o sistema ficaria mais sensível, mas ocorreu exatamente o contrário, gerando uma grande quantidade de falsos positivos e frequentemente não encontrando o nódulo. Por isso o músculo do peito é removido apenas após o processo de associação. Provavelmente isto ocorre devido à falta de

referência do músculo do peito, que contêm características semelhantes aos das regiões de massa, para os algoritmos de agrupamentos.

Um ponto a ser analisado é em relação aos tamanhos dos filtros. Em trabalhos como (AMARI; KASABOV, 1997; LEE, 2003; CADIEU et al., 2007), os autores comentam que os campos receptivos crescem consideravelmente cada vez que é necessário subir uma camada do córtex visual. Um estudo futuro deverá determinar um banco misto de funções bases, com diferentes tamanhos, a fim de obter uma melhor taxa de VP e menor taxa de FP na sugestão de diagnóstico de nódulos mamários.

Os resultados apresentados demonstram que a análise de componentes independentes realiza com êxito a extração de características, inspirada no conceito de codificação eficiente, para discriminar tecidos em imagens mamográficas. Além disso, observou-se que o *k-means* com as bases de ICA demonstrou um elevado desempenho preditivo para alguns conjuntos de dados e, assim, remeteu numa contribuição significativa para uma investigação clínica mais detalhada.

# 5. MÉTODO E RESULTADOS DA CLASSIFICAÇÃO

Neste capítulo é proposta uma metodologia para a classificação de regiões de interesse em mamografias digitalizadas, a partir de imagens bases geradas pelas técnicas de extração de características: *wavelets* de Gabor, análise de componentes principais e análise de componentes independentes. Para a classificação utiliza-se a técnica de análise discriminante linear. Nestas ROIs pode haver nódulos benignos, malignos ou tecidos normais, então esta metodologia serviria como um processo após a segmentação, apresentada no capítulo anterior, resultando em uma sugestão de diagnóstico para os radiologistas.

## 5.1. Introdução

Este método tem a intenção de realizar dois tipos de classificação dos tecidos mamários em mamografias digitais:

- classificação em nódulos ou não-nódulos;
- classificação em benignos ou malignos.

O método é baseado em três passos: (1) aquisição de imagens, (2) extração de características e (3) classificação. Todos estes passos estão ilustrados na Figura 26 e serão descritos neste capítulo.

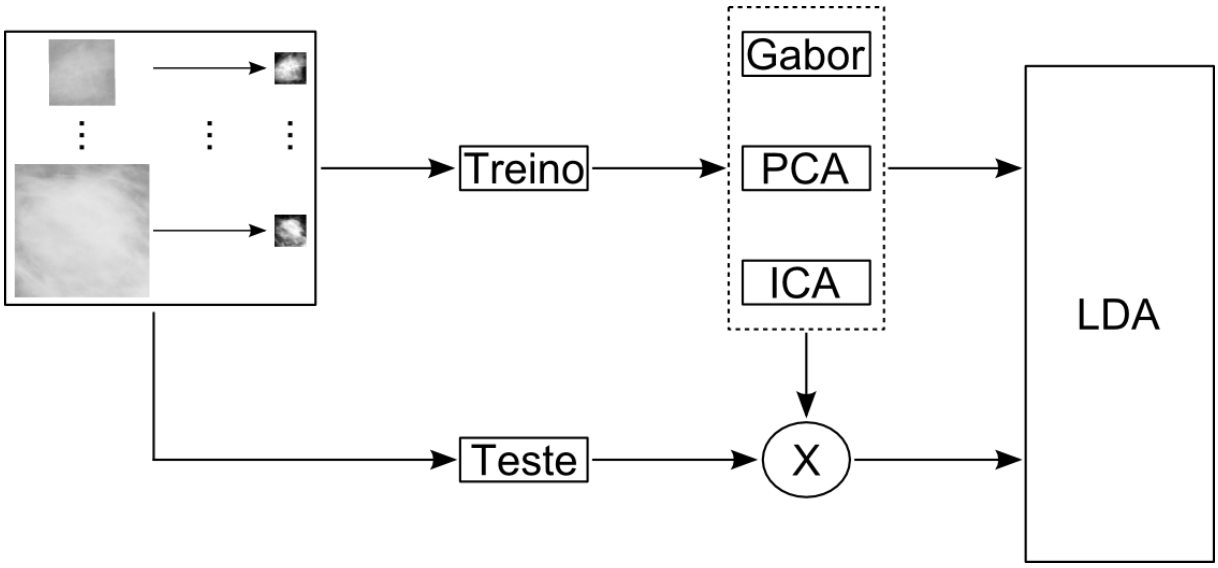


**Figura 26: Metodologia proposta em três passos para classificação de câncer de mama em imagens digitais. O primeiro passo é a aquisição de imagens, onde posteriormente serão divididas em dois grupos, treino e teste; a segunda etapa é a extração de características pelas técnicas de codificação: *wavelets* de Gabor, análise de componentes principais e análise de componentes independentes; o último passo é o de classificação pela análise discriminante linear, que dirá a que classe pertence o tecido analisado.**

### 5.2. Aquisição de Imagens

A aquisição da imagem é o passo onde é realizada a obtenção das mamografias e a seleção das regiões de interesse correspondentes ao tecido com e sem nódulo. A equação 49 (GONZALES; WOODS; EDDINS, 2009):

$$s_k = T(r_k) = \sum_{j=1}^{k} \frac{n_j}{n}, \tag{49}$$

para $k = 1,2,\cdots,L$, onde $s_k$ é o valor da intensidade do *pixel* $j$ da imagem processada, $n_j$ é o número de pixels com intensidade $r_k$ e $n$ é o total de números de *pixels*, foi utilizada para uma equalização do histograma, para enfatizar características não visíveis anteriormente nas ROIs. A Figura 27 ilustra uma equalização do histograma.

Em seguida as ROIs foram redimensionadas para $32 \times 32$ pixels, para que todas tenham o mesmo tamanho. Ao fim deste processo separam-se as imagens em dois grupos, $\mathbf{X}_{treino}$ e $\mathbf{X}_{teste}$, que são os conjuntos de treino e teste respectivamente.

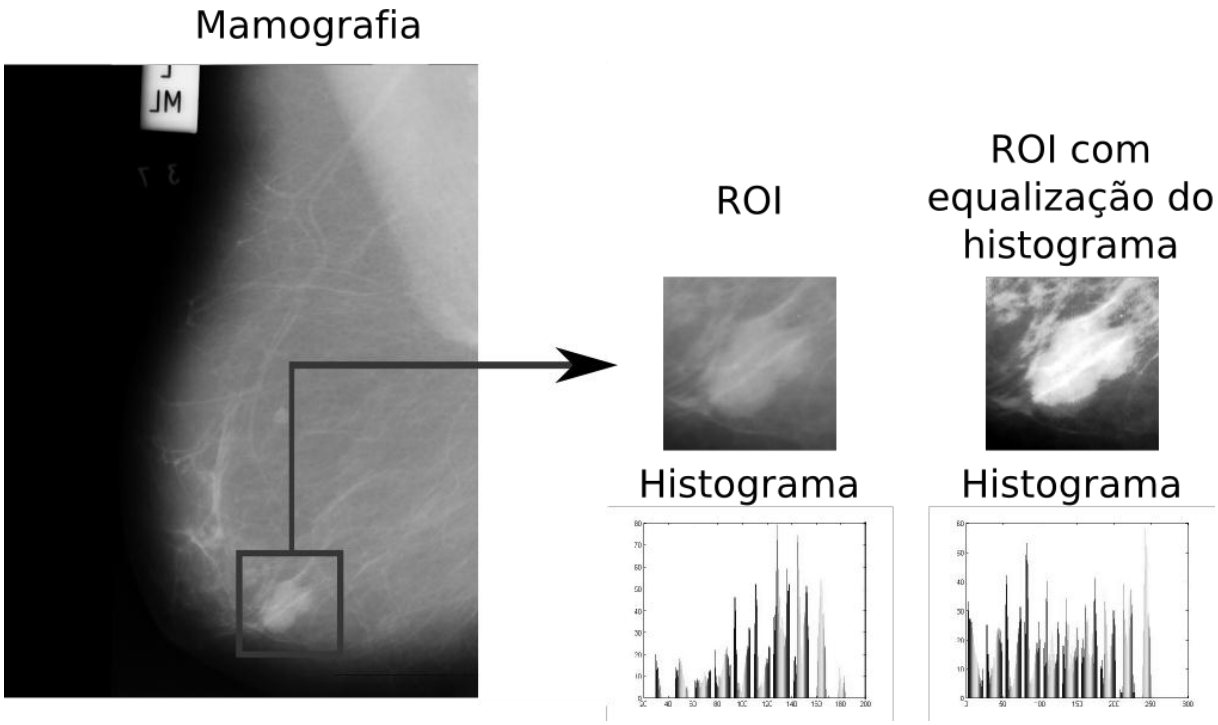


**Figura 27: Equalização do histograma de uma região de interesse. Observa-se que houve um realce na imagem após o procedimento de equalização do histograma. Fonte: Imagem editada a partir da imagem mdb005.pgm do banco de dados MIAS (SUCKLING et. al., 1994).**

Para o desenvolvimento e avaliação da metodologia proposta foi usada uma base de dados de mamografias disponível publicamente: a *Digital Database for Screening Mammography* (DDSM) (HEATH et al., 1998; HEATH et al., 2001).

A base de dados DDSM contêm 2620 casos com dois tipos de incidências padrão (médio-lateral obliquo e craniocaudal) de ambas as mamas, adquiridas do *Massachusetts General Hospital, Wake Forest University School of Medicine, Sacred Heart Hospital* e *Washington University* na *St. Louis School of Medicine*. Os dados compreendem estudos de pacientes de diferentes etnias. A DDSM contém descrições das lesões mamográficas em termos de imagens mamográficas da *American College of Radiology* chamada de *Breast Imaging Reporting and Data System* (BI-RADS) (HEATH et al., 1998; HEATH et al., 2001). Os mamogramas da base de dados DDSM foram digitalizadas por diferentes scanners dependendo da fonte dos dados de cada instituição e tem resolução entre 42 a 50 microns.

### 5.3. Extração de Características

O segundo passo da metodologia é a extração de características. São utilizados três diferentes métodos de codificação: *wavelets* de Gabor, PCA e ICA.

Pelas equações 1 e 2 (seção 3.2.1), foram geradas 100 *wavelets* de Gabor, com dez diferentes orientações (0°, 18°, 36°, 54º, 72º, 90º, 108º, 126º, 144º e 162º) e dez frequências (1,25 Hz, 2,22 Hz, 3,19 Hz, 4,16 Hz, 5,13 Hz, 6,11 Hz, 7,08 Hz, 8,05 Hz, 9,02 Hz e 10 Hz) a uma taxa de amostragem de 20 Hz.

Devida a grande quantidade de filtros, uma redução de dimensionalidade nos dados se torna necessária. Foram selecionados os filtros mais significativos para a classificação usando uma técnica de busca muito similar a descrita por SOUSA et al. (2007). Este processo consiste dos seguintes passos:

**Passo 1:** Definir um subespaço vazio $\mathbf{\Psi}$;

**Passo 2:** Repetir o passo seguinte para $k = 1, 2, \cdots, n$, onde $n$ é a dimensão de $\mathbf{\Psi}$;

**Passo 3:** Usar as equações da LDA para classificar as imagens $\mathbf{X}_{treino}^{T}$ projetadas no subespaço composto por $[\mathbf{\Psi}; \mathbf{G}_k]$. Onde $\mathbf{G}_k$ é a $k$-ésima *wavelet* de Gabor;

**Passo 4:** Selecionar as imagens bases de acordo com o melhor resultado da classificação do conjunto de treinamento;

**Passo 5:** Mover $\mathbf{G}_k$ de $\mathbf{G}$ para $\mathbf{\Psi}$ de modo que $n = n - 1$;

**Passo 6:** Retornar ao passo 2 até que $\mathbf{\Psi}$ obtenha a dimensão desejada.

As 50 *wavelets* de Gabor mais relevantes para esta classificação estão ilustradas na Figura 28.

Em seguida, será realizada a projeção de ambos os grupos, $\mathbf{X}_{treino}$ e $\mathbf{X}_{teste}$, no espaço de características escolhido pelo algoritmo de busca. As projeções são dadas pelas equações:

$$\begin{aligned} \widehat{\mathbf{Y}}_{treino} &= \mathbf{X}_{treino}^{T} \cdot \mathbf{\Psi}\,, \\ \widehat{\mathbf{Y}}_{teste} &= \mathbf{X}_{teste}^{T} \cdot \mathbf{\Psi}\,. \end{aligned} \tag{50}$$

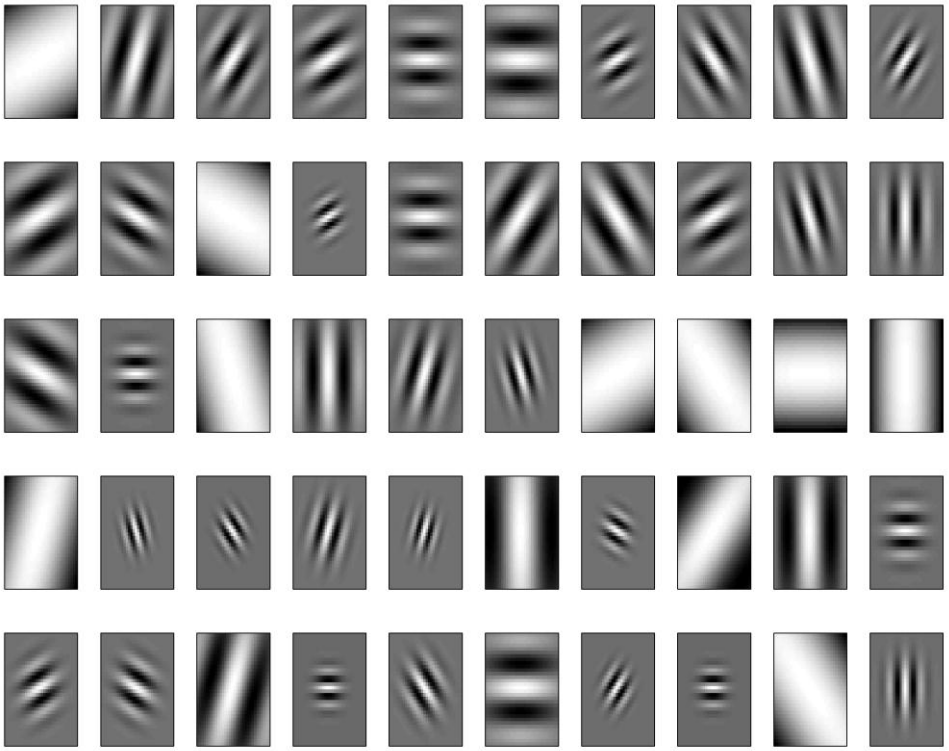

**Figura 28: *Wavelets* de Gabor escolhidas pelo algoritmo de busca. Dentre as 100 *wavelets* geradas, estas são as 50 mais significativas, conforme o algoritmo de busca desta metodologia.**

As características da PCA e da ICA são extraídas do conjunto de treinamento, conforme foi explicado na fundamentação teórica, capítulo 3. Tanto o processo de PCA quanto a técnica de ICA, geraram um total de 1024 filtros cada um. No caso da PCA, 1024 componentes principais e na de ICA 1024 funções bases. Devida a grande quantidade de características extraídas pelas técnicas, foi realizada uma redução de dimensionalidade utilizando os mesmos passos usados para selecionar as características mais relevantes para a classificação de Gabor. Substituindo-se apenas as *wavelets* pelas componentes principais no caso da PCA e pelas funções bases no caso da ICA. O algoritmo de busca para a PCA ficou da seguinte forma:

**Passo 1:** Definir um subespaço vazio $\boldsymbol{\Psi}$;

**Passo 2:** Repetir o passo seguinte para $k = 1, 2, \cdots, n$, onde $n$ é a dimensão de $\boldsymbol{\Psi}$;

**Passo 3:** Usar as equações da LDA para classificar as imagens $\mathbf{X}_{treino}^{T}$ projetadas no subespaço composto por $[\mathbf{\Psi}; \mathbf{V}_k]$. Onde $\mathbf{V}_k$ é a $k$-ésima componente principal;

**Passo 4:** Selecionar as imagens bases de acordo com o melhor resultado da classificação do conjunto de treinamento;

**Passo 5:** Mover $\mathbf{V}_k$ de $\mathbf{V}$ para $\mathbf{\Psi}$ de modo que $n = n - 1$;

**Passo 6:** Retornar ao passo 2 até quer $\mathbf{\Psi}$ obter a dimensão desejada.

As 50 componentes principais mais relevantes estão ilustradas na Figura 29.

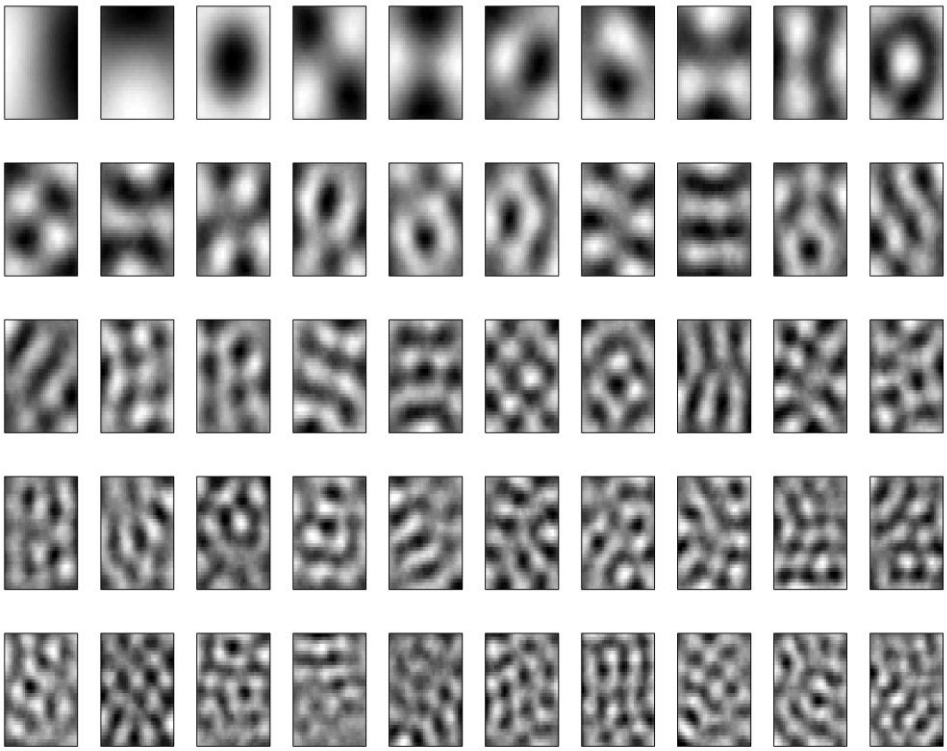

**Figura 29: Componentes principais escolhidas pelo algoritmo de busca. Dentre as 1024 componentes geradas, estas são as 50 mais significativas, conforme o algoritmo de busca desta metodologia.**

Em seguida, foi realizada a projeção de ambos os grupos, $\mathbf{X}_{treino}$ e $\mathbf{X}_{teste}$, no espaço de características escolhido pelo algoritmo de busca. Desta vez, o espaço $\mathbf{\Psi}$ está com as componentes principais, e não com as *wavelets* de Gabor como ocorreu no algoritmo anterior. As projeções são dadas pelas equações:

$$\begin{aligned} \hat{\mathbf{Z}}_{treino} &= \mathbf{X}_{treino}^{T} \cdot \boldsymbol{\Psi}\,, \\ \hat{\mathbf{Z}}_{teste} &= \mathbf{X}_{teste}^{T} \cdot \boldsymbol{\Psi}\,. \end{aligned} \tag{51}$$

Para a ICA, o algoritmo de busca ficou assim:

**Passo 1:** Definir um subespaço vazio $\boldsymbol{\Psi}$;

**Passo 2:** Repetir o passo seguinte para $k = 1, 2, \cdots, n$, onde $n$ é a dimensão de $\boldsymbol{\Psi}$;

**Passo 3:** Usar as equações da LDA para classificar as imagens $\mathbf{X}_{treino}^{\mathrm{T}}$ projetadas no subespaço composto por $[\boldsymbol{\Psi}; \boldsymbol{A}_{\mathrm{k}}]$. Onde, $\mathbf{A}_k$ é a $k$-ésima função base;

**Passo 4:** Selecionar as imagens bases de acordo com o melhor resultado da classificação do conjunto de treinamento;

**Passo 5:** Mover $\mathbf{A}_k$ de $\mathbf{A}$ para $\boldsymbol{\Psi}$ de modo que $n = n - 1$;

**Passo 6:** Retornar ao passo 2 até que $\boldsymbol{\Psi}$ obter a dimensão desejada;

As 50 funções bases mais relevantes estão ilustradas na Figura 30.

Como ocorreu com as *wavelets* e com a PCA, projetaram-se os grupos, $\mathbf{X}_{treino}$ e $\mathbf{X}_{teste}$, no espaço de características escolhido pelo algoritmo de busca. Onde $\boldsymbol{\Psi}$ representa as imagens bases geradas por este método de codificação eficiente. As projeções são dadas pelas equações:

$$\begin{aligned} \hat{\mathbf{X}}_{treino} &= \mathbf{X}_{treino}^{T} \cdot \boldsymbol{\Psi}\,, \\ \hat{\mathbf{X}}_{teste} &= \mathbf{X}_{teste}^{T} \cdot \boldsymbol{\Psi}\,. \end{aligned} \tag{1}$$

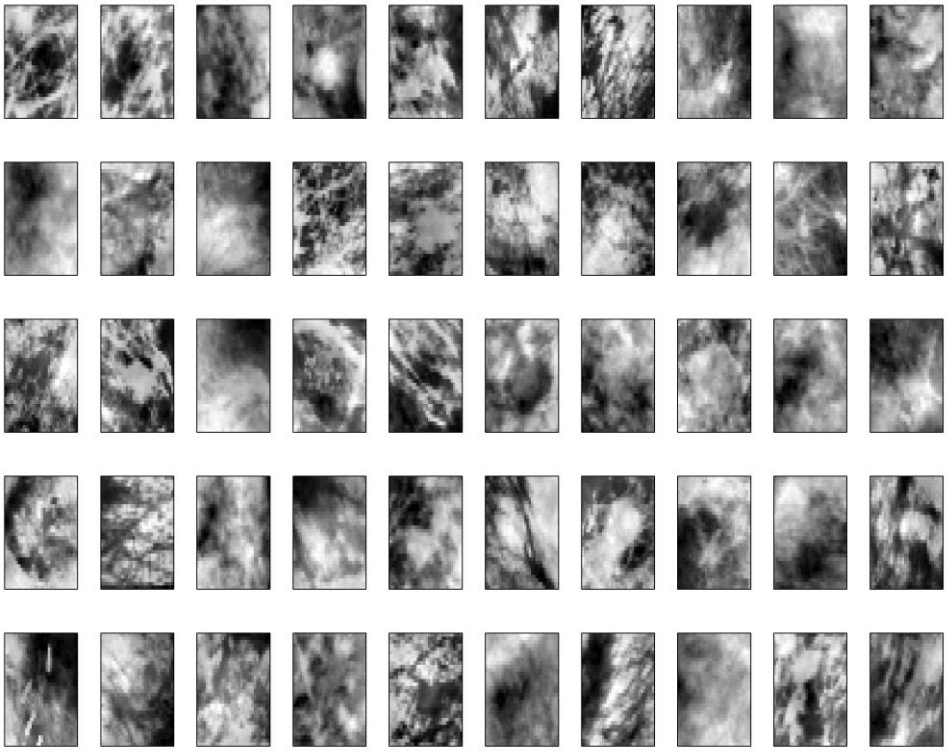

**Figura 30: 50 funções bases pelo modelo de codificação eficiente e selecionadas pelo algoritmo de busca.**

## 5.4. Classificação

No último passo, foi usada a análise discriminante linear (LDA), descrito na seção 3.4, para classificar os tecidos em nódulos ou não nódulos. Em casos de nódulos, classificou-os em benignos ou malignos.

### 5.5. Avaliação do Método

Para avaliar o classificador em relação à sua capacidade de diferenciação, analisa-se a sua sensibilidade, especificidade e acurácia. Para se entender melhor o que cada um destes termos significa, serão definidas as variáveis que serão utilizadas:

- Verdadeiro Positivo (VP) - Diagnóstico do nódulo classificado corretamente como um nódulo;
- Falso Positivo (FP) - Diagnóstico do não nódulo classificado erroneamente como um nódulo;
- Verdadeiro Negativo (VN) - Diagnóstico de não nódulo classificado corretamente como um não nódulo;
- Falso Negativo (FN) - Diagnóstico do nódulo classificado erroneamente como um não nódulo.

A acurácia é a taxa de sucesso ou acerto do teste e é dada por:

$$acurácia = \frac{(VP + VN)}{(VP + VN + FP + FN)}, \quad (2)$$

A sensibilidade indica o grau de qualidade do teste para identificar a patologia e é definida por:

$$sensibilidade = \frac{VP}{(VP + FN)}, \quad (3)$$

A especificidade indica o grau de qualidade do teste para identificar pacientes sem patologias e é definida por:

$$especificidade = \frac{VN}{(VN + FP)}. \quad (4)$$

O valor preditivo positivo (VPP) é a proporção de verdadeiros positivos entre todos os indivíduos com teste positivo. Expressa a probabilidade de um paciente com um teste positivo ter a doença. A VPP é definida por:

$$VPP = \frac{VP}{(VP + FP)}. \quad (5)$$

O valor preditivo negativo (VPN) é a proporção de verdadeiros negativos entre todos os indivíduos com o teste negativo. Expressa a probabilidade de um paciente com o teste negativo não ter a doença. A VPN é definida por:

$$VPN = \frac{VN}{(VN + FN)}. \quad (6)$$

Os cálculos para encontrar a acurácia, sensibilidade, especificidade, VPP e VPN na classificação entre benignos e malignos é dada de forma análoga.

### 5.6. Resultado

Dos 2620 casos do banco de dados DDSM, foram selecionados 3600 regiões de interesse, onde 900 são benignas, 900 são malignas e 1800 são normais. Todos os casos de regiões sem massa foram retirados de casos que não têm região de massa. Isto é, não foram retiradas regiões normais de casos que continham qualquer anormalidade.

Estas imagens foram divididas em dez grupos, para poder validar o sistema através do método *10-fold cross validation*, onde se utiliza um dos conjuntos para teste e todos os outros para o treinamento e repete-se o processo até que todos os grupos sejam testados. Desta forma os resultados aqui apresentados são a média dos dez testes realizados com este método.

O melhor resultado obtido com as *wavelets* de Gabor teve uma acurácia de 85,28% com 41 *wavelets*. A melhor precisão da PCA foi de 87,28% com 39 componentes principais e com a ICA obteve-se uma taxa de sucesso de 90,07% com 41 componentes independentes. A Figura 31 mostra a média dos resultados dada pelo método *10-fold cross validation* para

diferentes quantidades de componentes, de 1 à 50 componentes, para as três técnicas de extração de características aqui descritas. Analisando e comparando os resultados de precisão entre as técnicas, observa-se que a ICA atinge melhores resultados do que a PCA, a partir da primeira componente, e que os resultados, tanto na ICA quanto na PCA, ficam melhores com o aumento do número de componentes até o momento de convergência, em torno da quinta componente. A técnica de Gabor tem um ritmo diferente de aumento em relação à PCA e ICA. A referida técnica começa com uma queda em sua precisão até a 25ª componente e, em seguida, a taxa de sucesso aumenta até atingir seu pico em 41 componentes e a partir daí permanece estável.

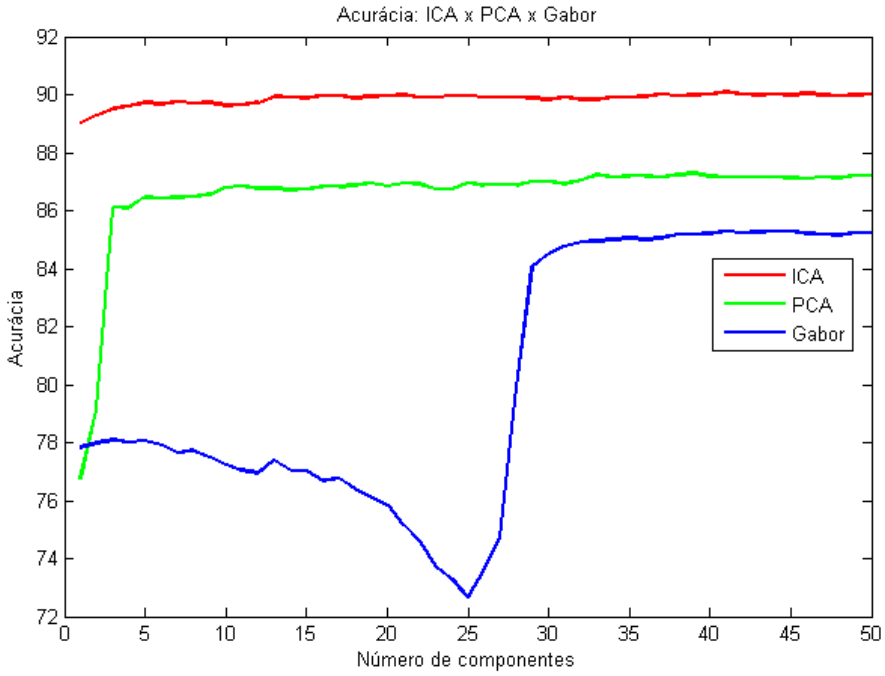


**Figura 31: Média das acurácias obtidas pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em massa ou não-massa.**

Em relação à sensibilidade e especificidade observou-se que o uso da técnica de ICA, obteve melhores resultados do que a PCA e Gabor (sensibilidade de 93,83% e especificidade de 85,48%, com 24 e 41 funções bases respectivamente), conforme Figuras 32 e 33, respectivamente. Com este resultado, observou-se que o sistema classifica os casos verdadeiros positivos melhor do que os casos verdadeiros negativos, garantindo boa confiabilidade aos casos clínicos.

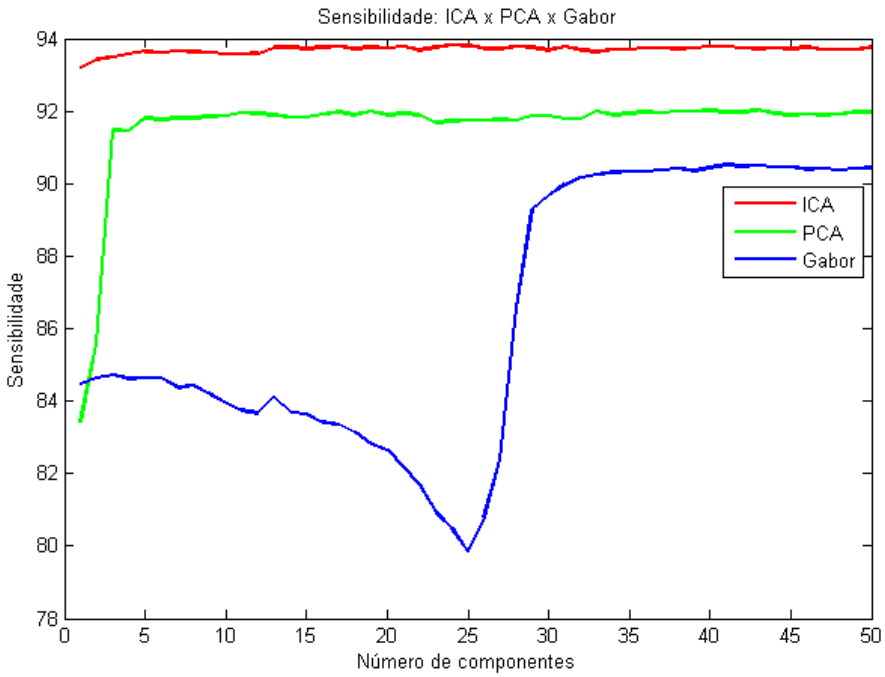


**Figura 32: Média das sensibilidades obtidas pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em massa ou não-massa.**

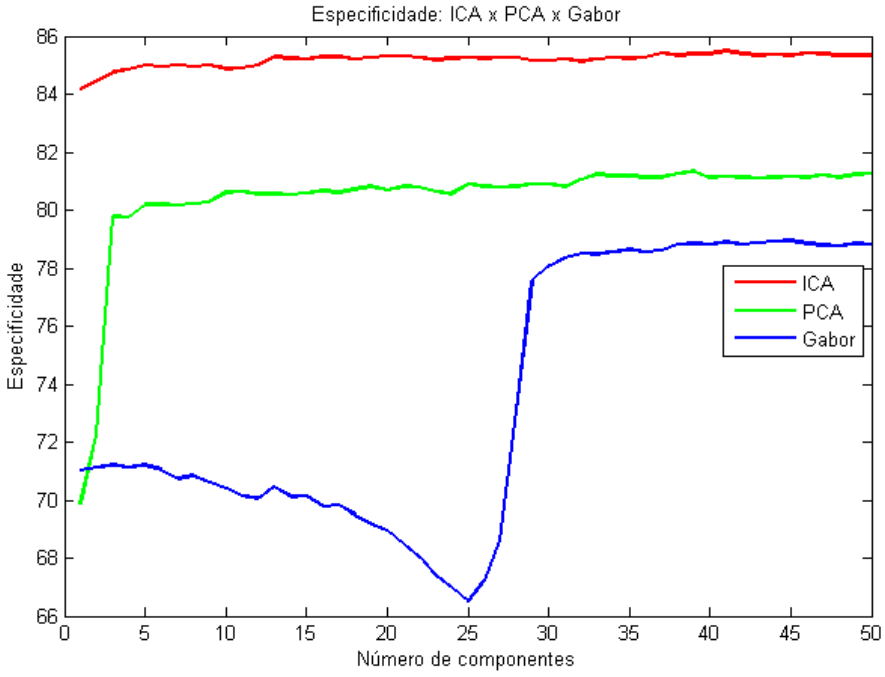


**Figura 33: Média das especificidades obtidas pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em massa ou não-massa.**

A ICA também foi a técnica que melhor obteve sucesso nas taxas de VPP e VPN, com 88,98% com 41 componentes e 91,69% com 24 componentes. Estes resultados são ilustrados nas figuras 34 e 35.

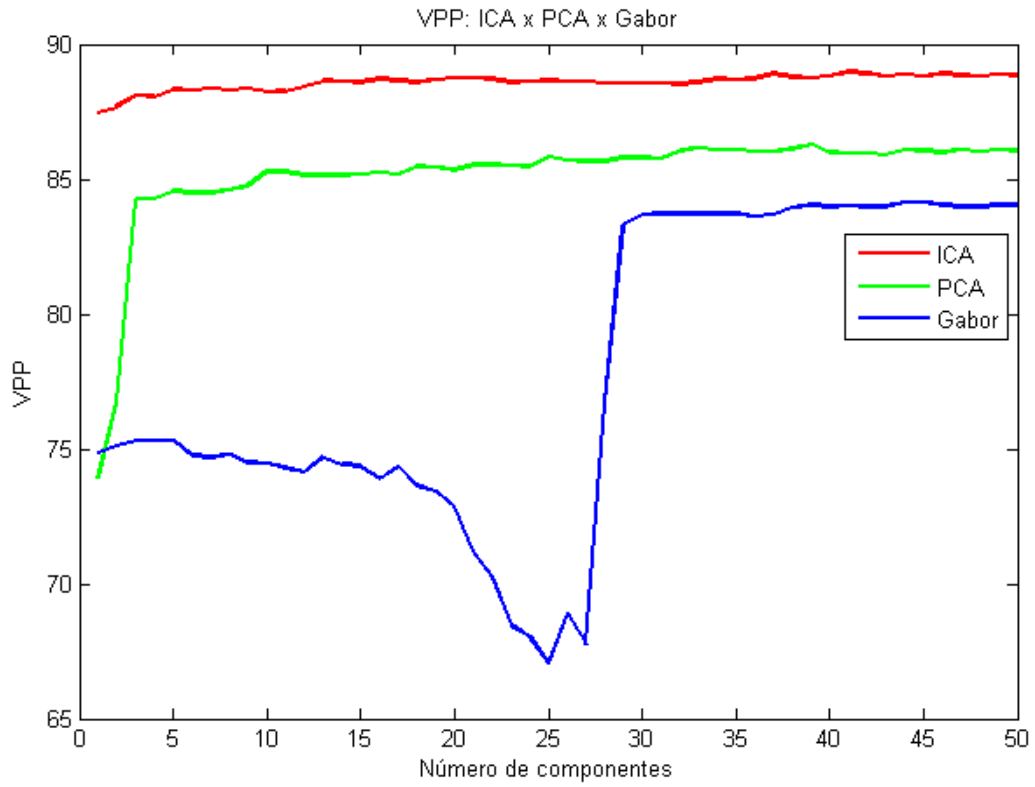


**Figura 34: Média dos valores preditivos positivos obtidos pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em massa ou não-massa.**

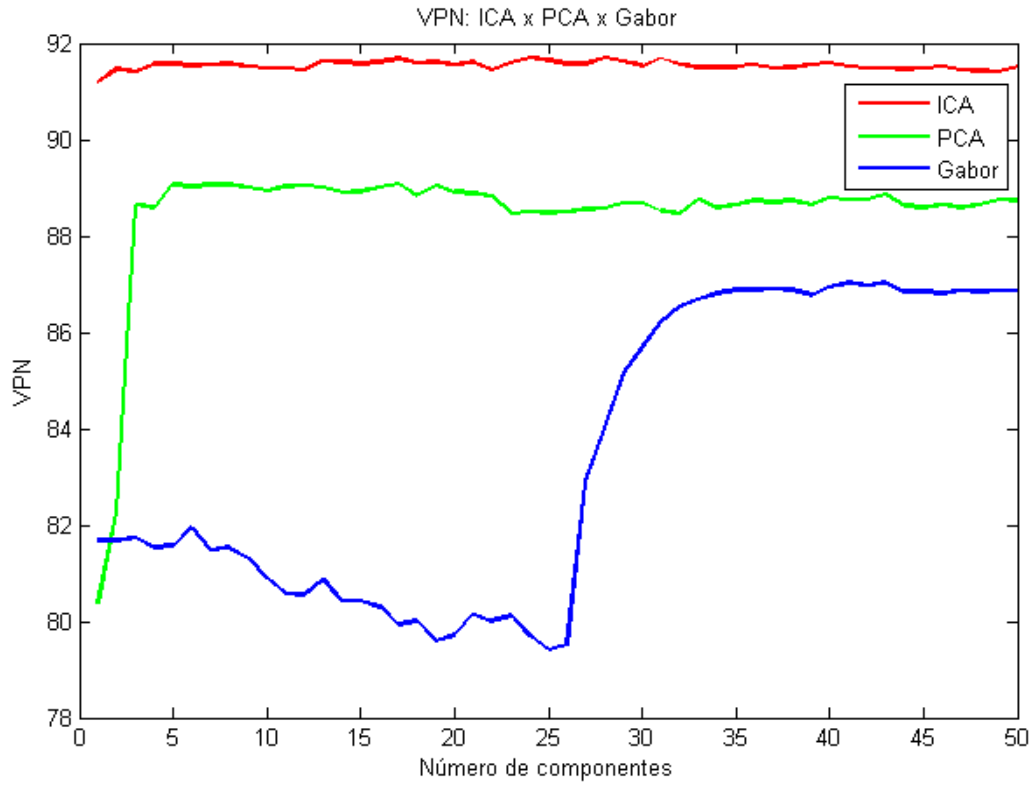


**Figura 35: Média dos valores preditivos negativos obtidos pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em massa ou não-massa.**

Outro estudo que foi conduzido está relacionado com os diagnósticos de nódulos como benigno ou maligno. Foram utilizados os mesmos bancos de filtros para obter a classificação de massa ou não-massa, e as mesmas técnicas para a classificação dos nódulos. Os resultados obtidos mostraram uma precisão média de 84,10%, com 18 componentes usando filtros de ICA. A sensibilidade foi de 82,84% com 12 componentes e a especificidade foi de 86,16% com sete componentes. A PCA resultou em uma precisão média de 81,48%, 79,91% de sensibilidade, ambas com três componentes e 84,06% de especificidade com nove componentes. O método Gabor resultou em uma taxa de sucesso de 77,00% e 75,95% de sensibilidade, ambas com 22 componentes, e 78,15% de especificidade com 11 componentes. Estes resultados podem ser observados nas Figuras 36, 37 e 38, respectivamente.

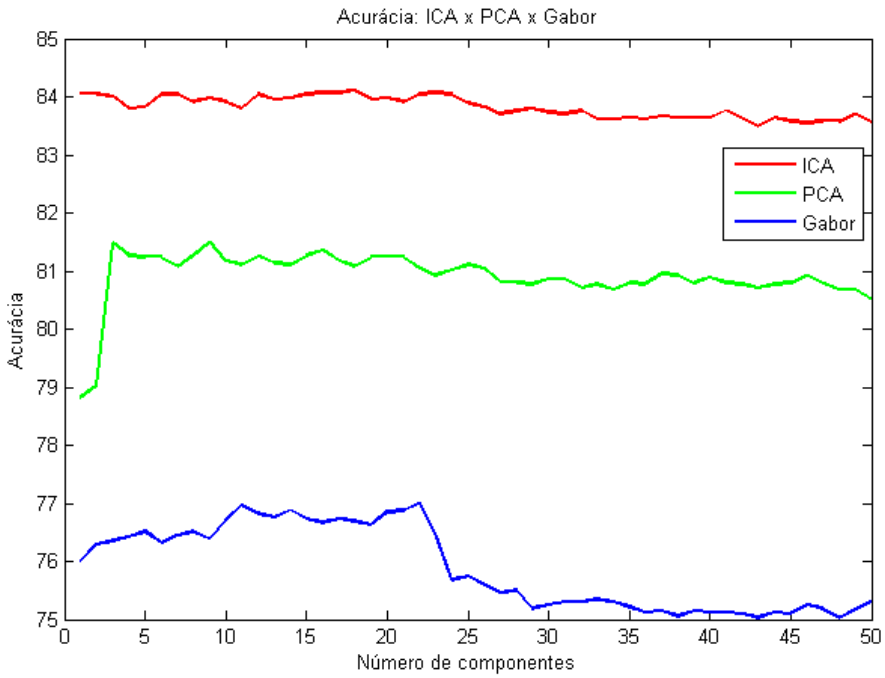


**Figura 36: Média das acurácias obtidas pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em benigno ou maligno.**

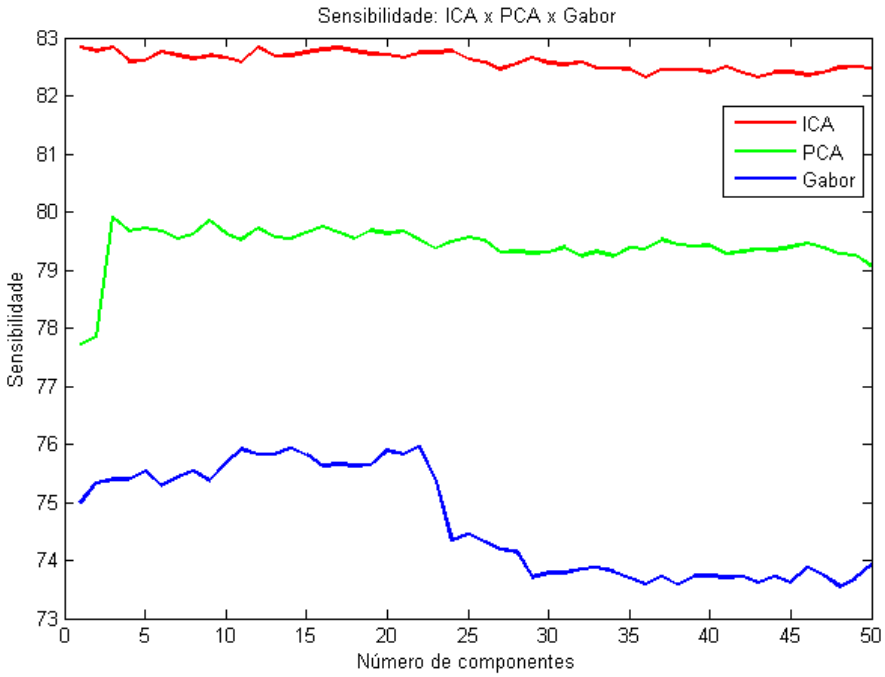


**Figura 37: Média das sensibilidades obtidas pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em benigno ou maligno**

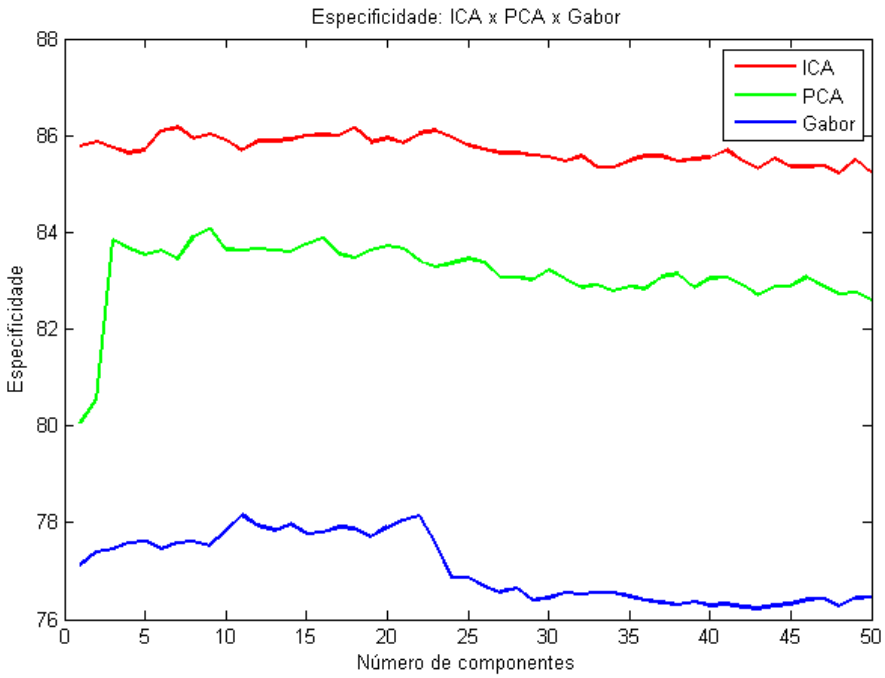


**Figura 38: Média das especificidades obtidas pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em benigno ou maligno**

A ICA novamente foi a técnica que melhor obteve sucesso nas taxas de VPP e VPN, com 90,30% com sete componentes e 77,61% com duas componentes. Estes resultados são ilustrados nas figuras 39 e 40.

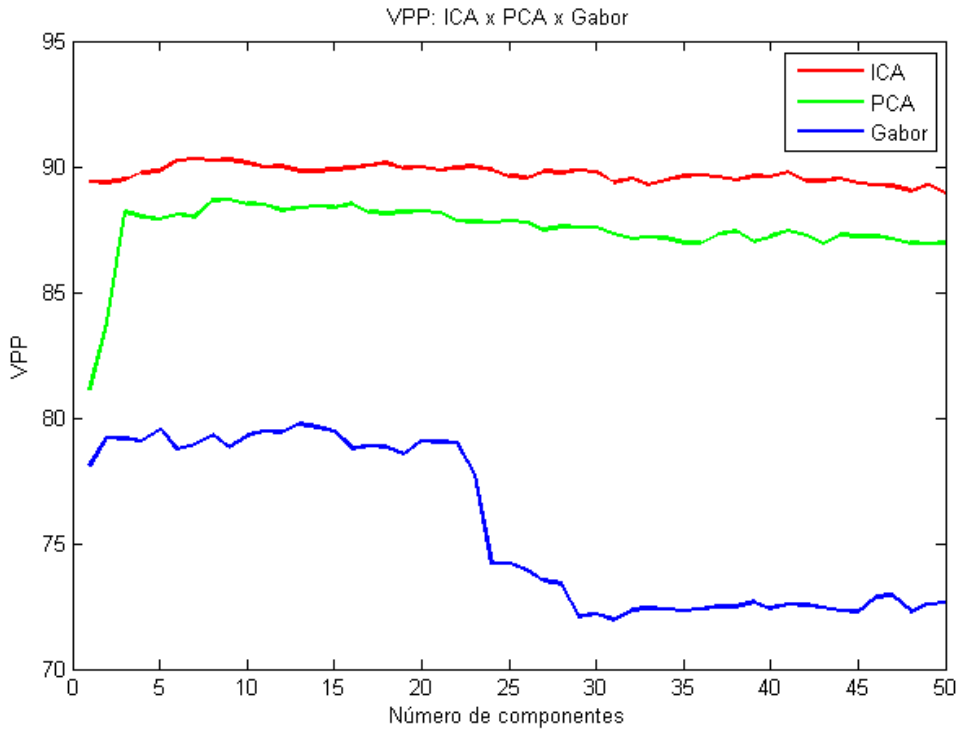


**Figura 39: Média dos VPP obtidos pelo método *10-fold cross validation* e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em benigno ou maligno.**

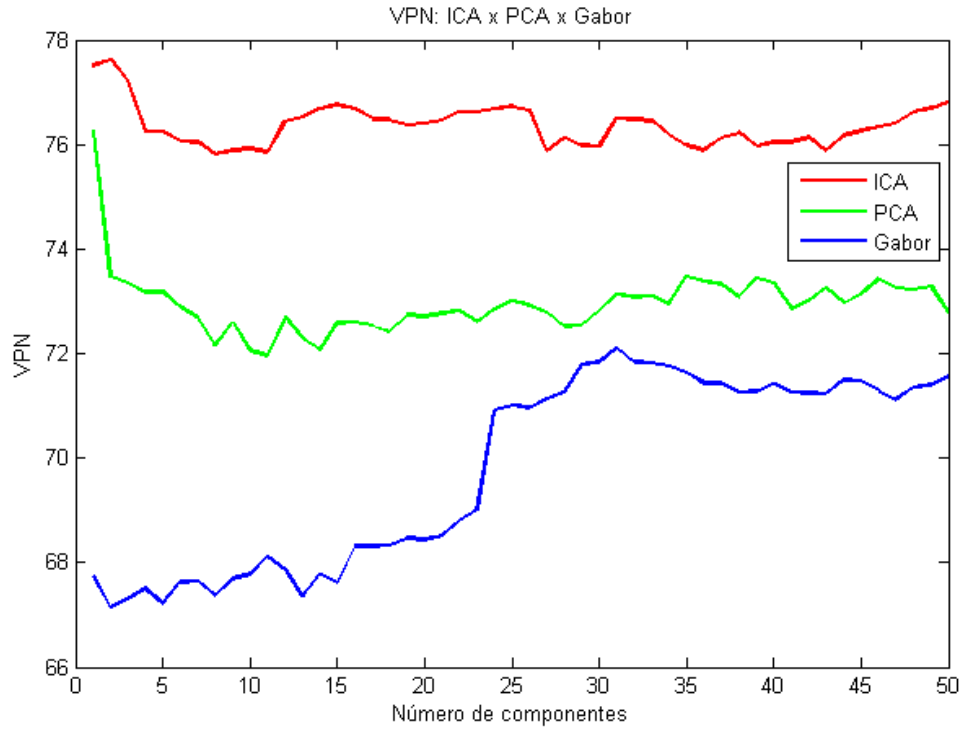


**Figura 40: Média dos VPN obtidos pelo método 10-fold cross validation e pelas três técnicas de codificação com diferentes números de componentes para a classificação de ROIs em benigno ou maligno.**

## 5.7. Discussões e Conclusões

Este capítulo apresentou um sistema de diagnóstico auxiliado por computador baseado no conceito de codificação eficiente, aplicado no problema de reconhecimento de nódulos mamários e classificando-os como massa ou não-massa. No caso de massa classifica-os como benigno ou maligno.

Apesar da diferença no resultado entre uma técnica e outra de codificação, principalmente entre PCA e ICA, ter sido pequena na taxa de sucesso, estes percentuais são muito valiosos, pois a ocorrência de falsos negativos (baixa sensibilidade) pode levar a paciente a óbito. A diferença da sensibilidade entre essas duas técnicas está em torno de 2%. Em números reais, os resultados obtidos neste trabalho são 72 pacientes melhor diagnosticadas. A especificidade entre PCA e ICA chegou a ser em torno de 4%, ou seja, 144 pacientes que não fizeram uma biópsia desnecessária, exame invasivo para a paciente e caro para o sistema de saúde.

A análise discriminante linear conseguiu separar parcialmente as duas classes, mas ainda existe uma margem de intersecção entre as mesmas, uma área que caracteriza os erros de classificação. Em trabalhos anteriores (COSTA et al., 2007), foi apresentado como resultado que o hiperplano gerado pela SVM separa melhor estas classes (algo em torno de 14% superior), proporcionando assim um resultado na classificação com maiores taxas de sucesso. No entanto, o custo computacional da LDA é menor do que o SVM, poupando tempo de operação. Também, comparando os resultados com outras técnicas que obtiveram uma maior acurácia, pode-se observar que houve uma única execução do teste nestes trabalhos, enquanto nesta tese utilizou-se da validação cruzada conhecida por *10-fold cross-validation*.

O número reduzido de imagens testadas também influencia no resultado, por exemplo, em (BRAZ et al., 2007) foram utilizadas 2048 imagens, 1024 para treino e 1024 para teste, (XU; PEI, 2011) aplicou seu método em 368 mamografias, (ISA; SIONG, 2012) empregou 322 mamogramas e (CAMPOS et al., 2005) usou apenas 200 imagens. Nesta metodologia foi testado um total de 3600 regiões de interesse, ou seja, o número de mamografias utilizadas foi de 3 a 18 vezes maior que os já utilizados. Estes trabalhos também não mencionam os

parâmetros necessários para um teste de significância, objetivando a comparação com os resultados obtidos neste trabalho.

Além disso, a suposição de linearidade conduz a uma limitação do nosso sistema, que não nos permite considerar estruturas não lineares na extração de características e classificação. Em trabalhos futuros, podem-se utilizar métodos não lineares de extração de características, tais como Kernel PCA (SCHÖLKOPF; SMOLA; MÜLLER, 1997), modelos ocultos não lineares de Markov (HMM, do inglês *Hidden Markov Model*) (WILSON, BOBICK, 1997) e outros modelos estatísticos (GHOSH; BOSE, 2005; SEIDE; MERTINS, 1994), a fim de conseguir uma possível melhora nas taxas de sucesso, visando um banco de imagens bases que possivelmente represente camadas superiores do córtex visual primário.

Um fator interessante a ser observado nos resultados é o fato de que a melhor precisão não é alcançada usando um maior conjunto de componentes, além disso, quanto maior a quantidade de componentes, mais a sensibilidade e a especificidade tendem para um valor menor. Suspeita-se que isso aconteça quando é usada muita informação redundante para classificar, confundindo o classificador, consequentemente diminuindo a taxa de sucesso. Acredita-se que o número ideal de componentes situa-se entre 30 a 50, até porque testes realizados com mais componentes não conseguiram melhores resultados, embora ainda tenha retornado resultados próximos da média.

Os resultados apresentados demonstram que a análise de componentes independentes realiza com êxito a extração de características, inspirada no conceito de codificação eficiente, para discriminar tecidos de massa e não-massa. Além disso, observou-se que a LDA com as bases de ICA demonstrou um elevado desempenho preditivo para alguns conjuntos de dados e, assim, remeteu numa contribuição significativa para uma investigação clínica mais detalhada.

# 6. CONSIDERAÇÕES FINAIS E PERSPECTIVAS FUTURAS

Este trabalho apresentou um sistema de diagnóstico auxiliado por computador baseado no conceito de codificação eficiente, aplicado no problema detecção e reconhecimento de nódulos mamários, classificando-os como massa ou não-massa. No caso de massa classifica-os como benigno ou maligno.

Devida a comparação realizada neste trabalho entre diferentes técnicas de codificação, observou-se que o fator principal para um diagnóstico preciso, tanto na segmentação quanto na classificação de tecidos mamários, é a extração de características. A taxa de sucesso está diretamente relacionada à escolha da técnica.

A análise de componentes independentes, baseada no conceito de codificação eficiente, realiza com sucesso esta extração de características. Pois os melhores resultados na segmentação e na classificação foram obtidos através dela. Acredita-se que este resultado foi melhor que o das outras técnicas porque além de garantir a descorrelação dos dados, garante-se também a independência estatística, representando exatamente o que Horace Barlow definiu em 1961 (BARLOW, 1961) como codificação eficiente: informação sem redundância através de suas componentes estatisticamente independentes.

Nos resultados, observou-se que tanto na segmentação quanto na classificação houve uma insensibilidade ao tamanho da massa, detectando regiões de diferentes tamanhos e posições. A síntese dos resultados desta tese pode ser observada na tabela 8.

**Tabela 8: Síntese dos resultados obtidos. As duas primeiras linhas são referentes à metodologia aplicada no capítulo 4, onde o segundo resultado foi obtido exclusivamente por mamas densas como teste. Os dois últimos resultados são referentes à metodologia do capítulo 5, onde o primeiro resultado diferencia regiões de massa das regiões de não-massa e o último é a discriminação entre imagens benignas de malignas.**

| Metodologia | Banco de imagens | Imagens | Segmentação | | Classificação (%) | | |
|---|---|---|---|---|---|---|---|
| | | | FLL | FNL | ACC | SEN | ESP |
| Capítulo 4 | MIAS | 118 | 0,96 | 2,13 | - | - | - |
| Capítulo 4 | MIAS (Densas) | 118 | 0,82 | 2,8 | - | - | - |
| Capítulo 5 | DDSM (massa/não-massa) | 3600 | - | - | 90,07 | 93,83 | 85,48 |
| Capítulo 5 | DDSM (benignas/malignas) | 1800 | - | - | 84,22 | 82,97 | 86,09 |

A partir dos resultados obtidos deste trabalho, gerou-se uma aplicação *web* para que o sistema possa ser utilizado por qualquer profissional cadastrado, Figura 41. Esta aplicação gerou uma dissertação de mestrado intitulada: "Implementação de um sistema de telediagnóstico para classificação de massas em imagens mamográficas usando análise de componentes independentes" (SILVA, 2012), onde foi proposta a modelagem e implementação de um sistema de telediagnóstico para análise e detecção automática de lesões em imagens mamográficas, baseado em análise de componentes independentes e máquina de vetor de suporte. O sistema analisa uma imagem de mamografia digital enviada pela internet e fornece o diagnóstico da imagem, indicando a presença de regiões suspeitas, que podem ser confirmadas por um especialista.

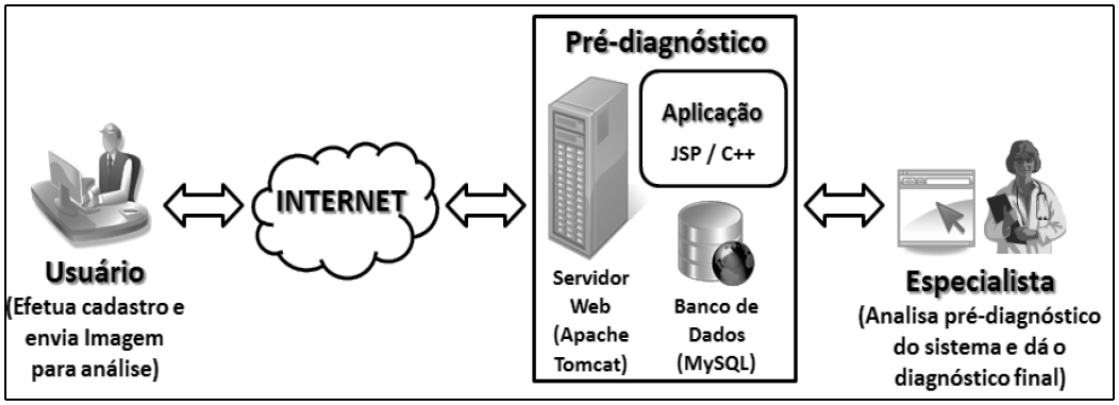


**Figura 41: Visão geral do sistema proposto por SILVA (2012). O usuário efetua o cadastro e envia a imagem através da internet para o servidor que irá realizar uma sugestão de diagnóstico e repassa a informação para um especialista emitir o diagnóstico final. Adaptada de (SILVA, 2012)**

Além dessa dissertação, este trabalho deu origem aos seguintes artigos:

- Costa, D. D.; Campos, L. F.; Barros, A. K. **Classification of breast tissue in mammograms using efficient coding.** BioMedical Engineering Online, v. 10, n. 55, 2011.

Este artigo descreve a metodologia aplicada nesta tese para a classificação de regiões de massa e não-massa através da técnica de codificação eficiente, como extrator de características, e da análise discriminante linear para a classificação.

- Costa, D. D.; Campos, L. F.; Barros, A. K. **Breast Image Classification Based on Texture Features Using Discriminant Analysis.** In: IEEE Conference Proceedings on Biosignals and Robotics for Better and Safer Living, Vitória: 2011.

Este trabalho propõe um método para a discriminação de tecidos mamográficos através das características de texturas da matriz de co-ocorrência de tons de cinza (GLCM) e da análise discriminante linear. Foi utilizada a base de imagens MIAS e como resultado foi obtido uma acurácia de 85,71%, sensibilidade de 80% e especificidade de 87,5%.

- Campos, L. F.; Lemos, E. C. M; Silva, L. C. O; Costa, D. D.; Barros, A. K. **Segmentation and Classification of Breast Cancer Using Independent Component Analysis, Texture Features and Neural Networks**. In: XI Workshop de Informática Médica (WIM 2011), Computação para Todos no Caminho da Evolução Social, Natal: 2011.

Foi proposto um método de segmentação e classificação de câncer de mama em mamografias digitais utilizando análise de componentes independentes (ICA), características de textura e redes neurais *perceptron* multicamadas (MLP). O método foi testado no conjunto de imagens mamográficas do MIAS, resultando na taxa de sucesso de 90,15%, com 92% de especificidade e 88,3% de sensibilidade.

- Campos, L. F. Costa, D. D.; Barros, A. K. **Detection of Breast Cancer in Digital Mammography using Independent Component Analysis and K-means**

**Clustering.** In: International Conference on Brain Inspired Cognitive Systems 2008 (BICS 2008), São Luís: 2008.

Neste trabalho foi proposto um método para segmentação de mamografias por meio de análise de componentes independentes e o algoritmo de agrupamento *k-means*. O método foi testado no banco de imagens mamográficas MIAS. O melhor desempenho foi obtido utilizando oito filtros de ICA, resultando em uma taxa de sucesso de 86,6%.

- Costa, D. D.; Campos, L. F.; Barros, A. K.; Silva, A. C. **Independent Component Analysis in Breast Tissues Mammograms Images Classification Using LDA and SVM.** In 6th International Special Topic Conference on Information Technology Applications in Biomedicine 2007 (ITAB 2007). Proceedings of the IEEE Engineering in Medicine and Biology Society, Tokyo: v. 6, p. 231-234, 2007.

Neste artigo, foi apresentada uma metodologia que utiliza análise de componentes independentes, juntamente com máquina de vetor de suporte (SVM) e análise discriminante linear (LDA) para distinguir entre tecidos de massa ou não-massa e tecidos benignos ou malignos de mamografias. Como resultado, encontramos o seguinte: LDA atingiu até 89,5% de precisão para discriminar massa ou não-massa e até 95,2% para discriminar benigno ou maligno no banco de dados DDSM. No banco de dados MIAS obteve-se até 85% de acurácia para discriminar tecidos de massa ou não-massa e até 88% entre benigno ou maligno. A SVM alcançou uma taxa de até 99,6% de precisão para discriminar massa ou não-massa e até 99,5% para discriminar tecidos benignos de Malignos no banco de dados DDSM. No banco de imagens MIAS obteve-se até 97% de sucesso na discriminação entre massa e não-massa e até 100% para discriminar benigno ou maligno.

- Guilhon, D.; Costa, D. D.; Van Leeuwen, P.; Hailer, B.; Barros, A. K.; Comani, S. **ICA-based pattern recognition system for the classification of Coronary Artery Disease patients studied with Magnetocardiography**. International Conference on Biomagnetism 2008 (Biomag 2008), Sapporo: 2008.

Neste trabalho, aplicou-se a análise de componentes independentes para reduzir a redundância nos MCGs registrados em 10 pacientes com indicação de angiografia coronária,

quatro dos quais foram realmente afetados pela doença arterial coronariana e seis não foram. O método de compressão baseado em ICA superou os sistemas baseados em PCA no sentido de alcançar altas taxas de compressão com baixa porcentagem de erro entre os sinais originais e reconstruídos. Erros na morfologia das ondas cardíacas foram insignificantes. As informações sobre as características dos sinais cardíacos fornecidos pela ICA foram então usadas para treinar uma rede neural *multilayer perceptron* (MLP) que automaticamente classifica os pacientes em portadores ou não da doença, com taxa de sucesso de 94,9%, especificidade de 95,7% e sensibilidade de 94,1%. Conclui-se que ICA pode ser eficientemente utilizado para comprimir dados de MCG, com escopo de facilitar e acelerar a manipulação de dados, e para apoiar o diagnóstico de pacientes com suspeita de doença arterial coronariana.

- Ribeiro, A. C.; Costa, D. D.; Barros, A. K.; Guilhon, D.; Comani,S; Braz Jr, G. **Diabetes Diagnosis Trough ICA and One Class SVM**. Brain Inspired Cognitive Systems 2008 (BICS 2008), São Luís: 2008.

Este estudo propõe um sistema de diagnóstico auxiliado por computador baseado em um método de codificação eficiente com análise de componentes independentes (ICA) e de máquina de vetor de suporte de uma classe (SVM) para classificar os dados de um conjunto de pacientes em diabéticos ou não diabéticos. Obteve-se um índice de precisão de diagnóstico de 99,84%, e este sistema alcançou um resultado muito promissor em relação a outras aplicações na literatura para este problema.

Como um trabalho futuro que deverá ser desenvolvido, pretende-se através da análise de componentes independentes, extrair informações de tecidos que possam desenvolver algum tipo de nódulo, ou seja, tentar prever o desenvolvimento de regiões de massa antes mesmo de sua formação. Um resultado parcial está ilustrado na Figura 42, onde se pode observar a mamografia em forma de um mapa de calor. Quanto mais vermelho, maiores as chances de estar desenvolvido ou estar se desenvolvendo algum tipo de patologia naquela região.

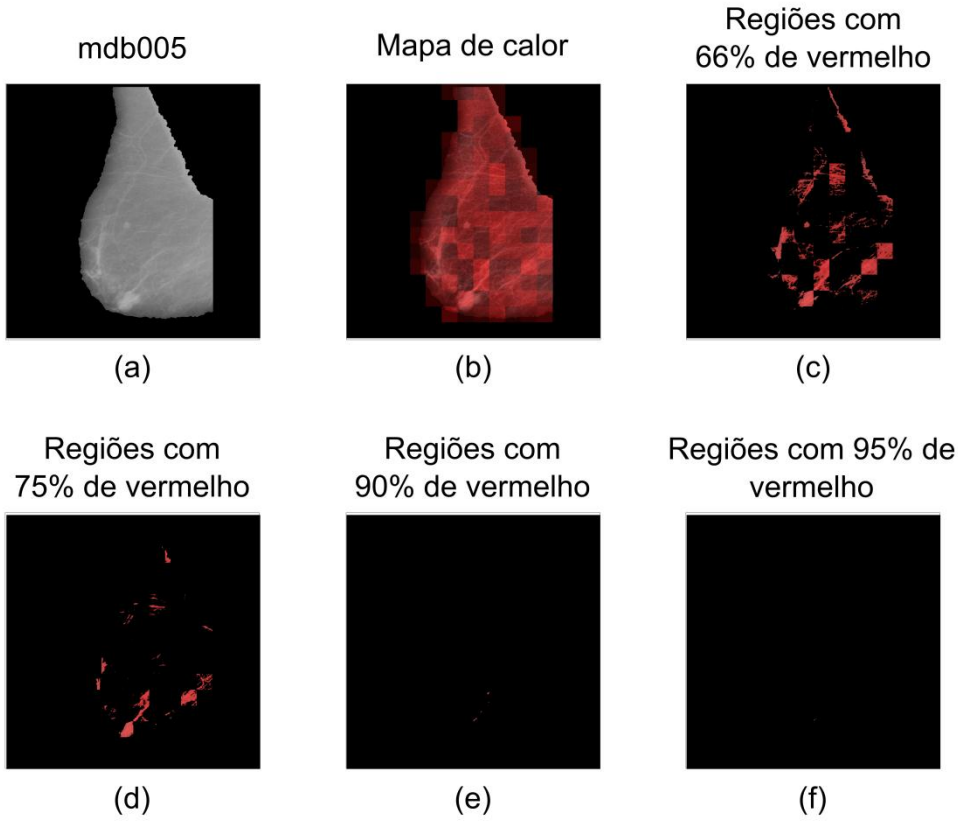


**Figura 42: Mapa de calor. Em (a) temos a imagem original, do banco de dados MIAS, da qual já foi retirado ruídos, artefatos e o músculo do peito. Em (b) temos a imagem original em forma de mapa de calor, quanto mais vermelho, maiores as chances daquela região conter algum tipo de nódulo. (c) São ilustradas apenas as regiões quentes que contenham pelo menos 66% de vermelho. Em (d) (e) e (f) são ilustradas respectivamente as regiões com 75%, 90% e 95% de vermelho.**

Percebe-se também que quanto maior for o limiar de vermelho selecionado para a ilustração da imagem, menor a quantidade de falsos positivos. Tanto que na Figura 38 (f), onde há regiões com pelo menos 95% de vermelho, apenas um ponto na vizinhança do nódulo é localizado.

Este método poderá ser validado com as imagens que serão adquiridas pelo sistema criado pelas metodologias aqui apresentadas. Enquanto o sistema vai sugerindo um diagnóstico ao especialista, ele também vai armazenando essas imagens para futuras análises. A comparação destas imagens antes e após o desenvolvimento de regiões de massa é que poderão realmente validar a metodologia. Sem contar a criação de um novo banco de imagens mamográficas com informações temporais.

As principais contribuições desta tese são:

- Avaliação de diferentes algoritmos de agrupamentos na detecção de nódulos mamários (*k-means*, nebuloso *c-means* e mapa auto-organizável);
- Avaliação de diferentes técnicas de extração de características (*wavelets* de Gabor, análise de componentes principais e análise de componentes independentes) para o reconhecimento de regiões de massa e classificação em mamografias digitalizadas.
- Avaliação do efeito da quantidade de funções bases utilizadas na classificação das regiões de interesse.
- Desenvolvimento, registro e implementação do protótipo do software de um sistema de diagnóstico auxiliado por computador para realização de sugestão de diagnósticos de câncer de mama, com intuito de ajudar os especialistas (protocolo do registro no INPIREMA: 14112012000073).

Os resultados demonstram que a análise de componentes independentes realiza com êxito a extração de características, inspirada no conceito de codificação eficiente, para discriminar tecidos de massa e não-massa. Além disso, observou-se que o algoritmo de *k-means* e LDA com as bases de ICA demonstraram um elevado desempenho preditivo para alguns conjuntos de dados e, assim, remeteu numa contribuição significativa para uma investigação clínica mais detalhada.

## REFERÊNCIAS


AMARI, S.; KASABOV, N. **Brain-Like Computing and Intelligent Information Systems**. 2. Ed. New York: Springer-Verlag, 1998.

AMERICAN CANCER SOCIETY. **Breast Cancer Facts & Figures 2011-2012**. Atlanta: American Cancer Society, 2011a. Disponível em: <http://www.cancer.org/Research/CancerFactsFigures/BreastCancerFactsFigures/ACSPC-030975>. Acessado em: 11/12/2012.

AMERICAN CANCER SOCIETY. **Cancer Facts & Figures 2007.** Atlanta: American Cancer Society, 2007.

AMERICAN CANCER SOCIETY. **Cancer Prevention & Early Detection Facts & Figures 2011.** Atlanta: American Cancer Society, 2011b.

AMERICAN CANCER SOCIETY. **Early Detection, Diagnosis and staging topics**. 2012. Disponível em: <http://www.cancer.org/Cancer/BreastCancer/DetailedGuide/breast-cancer-detection>. Acesso em: 30/10/2012.

BALLEYGUIER, C.; KINKEL, K.; FERMANIAN, J.; MALAN, S.; DJEN, G.; TAOUREL, P.; HELENON, O. **Computer-aided Detection (CAD) in Mammography: Does it Help the Junior or the Senior Radiologist?**. European Journal of Radiology, v. 1, p. 90-96, 2005.

BARLOW, H. **Possible principles underlying the transformations of sensory messages.** In: ROSENBLITH, W. **Sensory Communication.** Cambridge: MIT Press, 1961, cap. 13, p. 217-234.

BARLOW, W. E. et al. **Performance of Diagnostic Mammography for women with signs or symptoms of breast cancer.** Journal of the National Cancer Institute, v. 94, n. 15, p. 1151-1159, 2002.

BAUER, W.; IGOT, J. P.; LE, G.Y. **Chronology of breast cancer using Gompertz' growth model.** Annales d'anatomie pathologique. Paris: v. 25, nº 1, p. 39-56, 1980.

BISHOP, C. M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. 1. ed. New York: Springer Science + Business Media LLC, 2006.

BLAND, K. I.; COPELAND, E. M. **La Mama, manejo multidisciplinarios de lãs enfermedades benignas y malignas.** 3. ed. Buenos Aires: Médica Panamericanas, 2000, Cap. 13, p. 290-300, 2000.

BLAND, K. I.; COPELAND, E. M; **A mama: tratamento compreensivo das doenças benignas e malignas.** 1. ed. São Paulo: Manole, 1994.

BOVIC, A. C.; ACTON, S. T. Basic Linear Filtering with Application to Image Enhancement. In: BOVIK, A. **Handbook of Image & Video Processing**. 1. ed. San Diego: Academic Press, 2000.

BOYD, N. F.; BYNG, J. W.; JONG, R. A.; FISHELL, E. K.; LITTLE, L. E.; MILLER, A. B.; LOCKWOOD, G. A.; TRITCHELER, 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, nº 9, p. 670-675, 1995.

BRANDÃO, A. C. C. **Boas Práticas Farmacêuticas - Blog**. Disponível em: <http://boaspraticasfarmaceuticas.blogspot.com.br/2011/10/mamografia-e-um-grande-aliado-na-luta.html>. Acesso em: 07/08/2012.

BRAZ, JR.; SILVA, G, E. C.; PAIVA A. C; SILVA, A. C. **Breast Tissues Classification Based on the Application of Geostatistical Features and Wavelet Transform.** In: 6th International Special Topic Conference on Information Technology Applications in Biomedicine (ITAB 2007), v. 6, p. 227-230, Tokyo, 2007.

CADIEU, C., KOUH, M., PASUPATHY, A., CONNOR, C. E., RIESENHUBER, M., POGGIO, T. **A Model of V4 Shape Selectivity and Invariance.** Journal of Neurophysiology, v. 98, nº 3, p. 1733–1750, 2007.

CALAS, M. J. G; GUTFILEN, B; PEREIRA, W. C. A. **CAD e mamografia: por que usar esta ferramenta?** Radiologia Brasileira, v. 45, nº 1, 2012.

CAMPANINI, R.; BAZZANI, A.; BEVILACQUA, A.; BOLLINI, D.; et al. **A novel approach to mass detection in digital mammography based on Support Vector Machines (SVM).** In: Proceedings of the 6th International Workshop in Digital Mammography (IWDM). Springer Verlag, p. 399-401, Bremem: 2002.

CAMPOS, L. F. A.; SILVA, A. C.; BARROS, A. K. **Diagnosis of Breast cancer in digital mammograms using independent component analysis and neural networks.** Lecture Notes in Computer Science (LNCS). Berlin: v. 3773, p. 460-469, 2005.

CAMPOS, L. F. COSTA, D. D.; BARROS, A. K. **Detection of Breast Cancer in Digital Mammography using Independent Component Analysis and K-means Clustering**. In: Brain Inspired Cognitive Systems 2008 (BICS 2008), São Luís, 2008.

CARANDINI, M.; HEEGER, D. J.; MOVSHON, J. A. **Linearity and Normalization in Simple Cells of the Macaque Primary Visual Cortex**. The Journal of Neuroscience, v. 17, n. 21, p. 8621-8644, 1997.

CASTELLS, F; LAGUNA, P; SÖRNMO, L; BOLLMANN, A; ROIG, J. M. **Principal Component Analysis in ECG Signal Processing**, Hindawi Publishing Corporation, EURASIP Journal on Advances in Signal Processing, v. 7, p. 1-21, 2007.

CAVALCANTE, A. B.; MANDIC, D. P.; RUTKOWSKI, T. M.; BARROS, A. K. **Speech Enhancement Based on the Response Features of Facilitaded EI Neurons.** In: Proceedings of Independent Component Analysis and Blind Signal Separation, 6th International Conference (ICA 2006). Lecture Notes in Computer Science (LNCS), v. 3889, p. 585-592, 2006.

CHEN, Y.; CHANG, C. I. **A new application of texture unit coding to mass classification for mammograms.** In: International Conference on Image Processing (ICIP 2004), v. 5, p. 3335-3338, 2004.

CHRISTOYIANNI, I.; KOUTRAS, A.; DERMATAS, E.; KOKKINAKIS, G. **Computer aided diagnosis of breast cancer in digitized mammograms.** Computerized Medical Imaging and Graphics, v. 26, p. 309-319, 2002.

CONSELHO REGIONAL DE MEDICINA DO ESTADO DO MATO GROSSO (CRM-MT). **SUS não faz biópsia de mama há 6 meses**. Disponível em: <http://www.crmmt.cfm.org.br/index.php?option=com_content&view=article&id=21003:sus-nao-faz-biopsia-de-mama-ha-6-meses&catid=3>. Acesso em: 01/08/2012.

COSTA, D. D.; CAMPOS, L. F.; BARROS, A. K.; SILVA, A. C. **Independent Component Analysis in Breast Tissues Mammograms Images Classification Using LDA and SVM**. In 6th International Special Topic Conference on Information Technology Applications in Biomedicine 2007 (ITAB 2007). Proceedings of the IEEE Engineering in Medicine and Biology Society, Tokyo: v. 6, p. 231-234, 2007.

COSTA, M. P.; CASTIER, M. B. **Alterações da função distólica como efeito do tratamento do câncer de mama.** Revista Hospital Universitário Pedro Ernesto, v. 11, n. 1, p. 65-70, 2012.

COVER, T; THOMAS, J. **Elements of Information Theory.** 2. ed. New York: Wiley, 1991.

DAYAN, P.; ABBOTT, L. F. **Theoretical Neuroscience.** 1. ed. Cambridge: MIT Press, 2001.

DE VALOIS, R. L.; ALBRECHT, D. G.; THORELL, L. G. **Spatial frequency selectivity of cells in macaque visual cortex**. Vision Research., v. 22, p. 545–559, 1982.

DOMÍNGUEZ, A. R.; NANDI, A. K. **Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection.** Computerized Medical Imaging and Graphics, v. 32, n. 4, p. 304-315, 2008.

DUNN, J. C. **A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters.** Journal of Cybernetics and Systems, v. 3, n. 3, p. 32-57, 1973.

ELTONSY, N. H.; TOURASSI, G. D.; ELMAGHRABY, A. S. **A Concentric Morphology Model for the Detection of Masses in Mammography**. IEEE Transactions on Medical Imaging, v. 26, n. 6, p. 880-890, 2007.

FERNANDES, R. A. Q.; NARCHI, N. Z. **Enfermagem e saúde da mulher.** 1. ed. São Paulo: Editora Manole, 2007.

FILHO, O. M.; NETO, H. V. **Processamento Digital de Imagens.** 1. ed. Rio de Janeiro: Brassport, 1999.

GAN, G.; MA, C.; WU, J. **Data Clustering: Theory, Algorithms, and Applications**. 1. ed. Philadelphia: ASA-SIAM Series on Statistics and Applied Probability, 2007.

GHOSH, A. K.; BOSE, S. **Feature Extraction for Nonlinear Classification.** In: Proceedings of the First International Conference of Pattern Recognition and Machine Intelligence (PReMI'05), Lecture Notes in Computer Science (LNCS), v. 3776, p. 170-175, 2005.

GONZALES, R. C.; WOODS, R. E. **Processamento de imagens digitais**. 3. ed. Tradução Cristina Yamagami e Leonardo Piamonte; Revisão técnica Marcelo Andrade da Costa Vieira e Mauricio Cunha Escarpinati. São Paulo: Pearson, 2010.

GONZALES, R. C.; WOODS, R. E.; EDDINS, S. L. **Digital Image Processing Using Matlab**. 2. ed. New York: Gatesmark, 2009.

GRAM, I. T.; LUND, E.; SLENKER, S. E. **Quality of life following a false positive mammogram**. Breast Journal Cancer, v. 62, p. 1018-1022, 1990.

GUILHON, D.; BARROS, A. K.; COMANI, S. **ECG Compression by Efficient Coding.** In: Proceedings of the 7th international conference on Independent Component Analysis and Signal Separation, Lecture Notes in Computer Science (LNCS), v. 4666, p. 593-600, 2007.

HARVARD MEDICAL SCHOOL - PORTUGAL PROGRAM. **Mamografia.** Revisado por Carvalho, A. e validado por Pereira, T. Disponível em: <http://mednet.umic.pt/portal/server.pt/community/Procedimentos/Procedimentos$Detail?idProcedimentos=AZP0033_006>. Acesso em: 07/08/2012.

HATHAWAY, R. J.; BEZDEK, J. C. **Recent Convergence Results for the Fuzzy c-Means Clustering Algorithms**. Journal of Classification, v. 5, p. 237-247, 1988.

HAYKIN, S. **Neural Networks and Learning Machines**. 3. ed. New Jersey: Pearson Prentice Hall, 2009.

HEATH, M; BOWYER, K; KOPANS, D; KEGELMEYER, W. P.; MOORE, R; CHANG, K.; MUNISHKUMARAN, S. **Current status of the Digital Database for Screening Mammography**. In: Proceedings of the Fourth International Workshop on Digital Mammography, Kluwer Academic Publishers, p. 457-460, 1998.

HEATH, M; BOWYER, K; KOPANS, D; MOORE, R; KEGELMEYER, W. P. **The Digital Database for Screening Mammography**. In: Proceedings of the Fifth International Workshop on Digital Mammography, M.J. Yaffe, ed., Medical Physics Publishing, p. 212-218, 2001.

HUBEL, D. H.; WIESEL, T. N. **Receptive fields and functional architecture of monkey striate cortex**. Journal of Physiology, v. 195, p. 215–243, 1968.

HYVÄRINEN, A. **Fast and Robust Fixed-Point Algorith for Independent Component Analysis**. IEEE Transaction on Neural Network, v. 10, nº 3, p. 626-634, 1999

HYVÄRINEN, A.; KARHUNEN, J.; OJA, E. **Independent Component Analysis**, Nova York: John Wiley & Sons, 2001..

ISA, N. A. M.; SIONG, T. S. **Automatic Segmentation and Detection of Mass in Digital Mammograms.** In: Proceedings of the 11th international conference on Telecommunications

and Informatics, Proceedings of the 11th international conference on Signal Processing, p. 143-146, 2012

JAIN, A. K. **Fundamentals of Digital Image Processing**. 1. ed. Englewood Cliffs: Prentice Hall, 1989.

JIN, J.; WANG, X.; WANG, B. **Classification of Direction perception EEG Based on PCA-SVM**. In: Third International Conference on Natural Computation 2007 (ICNC 2007), v. 2, p. 116-120, 2007.

JONES, J. P.; PALMER, L. A. **An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex**. Journal of Neurophysiology, v. 58, n. 6, p. 1233–1258, 1987.

KARKLIN, Y; SIMONCELLI, E. P. **Efficient coding of natural images with a population of noisy Linear-Nonlinear neurons.** In: Advances in Neural Information Processing Systems (NIPS'11), v. 24, p. 999-1007, 2011.

KOHONEN, T. **The self-organizing map**. In: Proceedings of the IEEE, v. 78, n. 9, p. 1464-1480, 1990.

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

LACHENBRUCH, P.A., **Discriminant Analysis.** New York: Hafner Press, 1975.

LEE, T. S. **Computations in the early visual cortex**. Journal of Physiology, v. 97, p. 121–139, 2003.

LENNIE, P. **Receptive Fields.** Current Biology, v. 13, n. 6, p. 216-219, 2003.

LEWICKI, M. S. **Efficient coding of natural sounds.** Nature Neuroscience, v. 5, n. 4, p. 356-363, 2002.

LIM, W. K.; ER, M. J. **Classification of Mammographic Masses using Generalized Dynamic Fuzzy Neural Networks.** Medical Physics, v. 31, n. 5, p. 1288-1295, 2004.

MARTINS, L. O.; SANTOS, A. M.; SILVA, A. C.; PAIVA, A. C. **Classification of Normal, Benign and Malignant Tissues using Co-Occurrence Matrix and Bayesian Neural Network in Mammographic Images.** In: Proceedings of the Ninth Brazilian Symposium on Neural Networks (SBRN'06), p. 24-29, 2006.

MARTINS, L. O.; SILVA, A. C.; PAIVA, A. C.; GATTASS, M. **Detection of Breast Masses in Mammogram Images Using Growing Neural Gas Algorithm and Ripley's K Function**. Journal of Signal Process System, v. 55, p. 77-90, 2009.

MASALA, G. L. **Computer Aided Detection on Mammography**. In: World Academy of Science, Engineering and Technology (WASET 2006), v. 15, p. 1-6, 2006.

MASSOTTI, M. **A Ranklet-Based Image Representation for Mass Classification in Digital Mammograms.** Medical Physics, v. 33, nº 10, p. 3951-3961, 2006.

MILLER, A. B. **Practical Applications for Clinical Breast Examination (CBE) and Breast Self-Examination (BSE) in Screening and Early Detection of Breast Cancer**. Breast Care (Basel), v. 3, n. 1, p. 17-20, 2008.

MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER. **Estimativa 2012: incidência de câncer no Brasil.** Rio de Janeiro: INCA, 2011.

MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER. **Mamografia: da prática ao controle – Recomendações para profissionais de saúde**. Rio de Janeiro: INCA, 2007. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/qualidade_mamografia.pdf>. Acessado em: 11/12/2012.

MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER. **Novidade na estimativa do câncer.** Rio de Janeiro: INCA, 2012.

MINISTÉRIO DA SAÚDE. INSTITUTO NACIONAL DO CÂNCER. **Parâmetros para o rastreamento do câncer de mama: recomendações para gestores estaduais e municipais**. Rio de Janeiro: INCA, 2009.

MINISTÉRIO DA SAÚDE. SECRETARIA DE ATENÇÃO A SAÚDE. INSTITUTO NACIONAL DO CÂNCER. **TNM: Classificação de Tumores Malignos**. Tradução Ana Lúcia Amaral Eisenberg. 6. ed. Rio de Janeiro: INCA, 2004a.

MINISTÉRIO DA SAÚDE. SECRETARIA DE ATENÇÃO A SAÚDE. INSTITUTO NACIONAL DO CÂNCER. **Controle do câncer de mama – Documento de conseso.** Rio de Janeiro: INCA, 2004b.

MINISTÉRIO DA SAÚDE. SECRETARIA DE ATENÇÃO A SAÚDE. INSTITUTO NACIONAL DO CÂNCER. **Detecção precoce ajuda tratamento contra câncer de mama.** Rio de Janeiro: INCA, 2007. Disponível em: <http://www.INCA.gov.br/releases/press_release_view.asp?ID=1355>. Acesso em 31/01/2011.

MINISTÉRIO DA SAÚDE. SECRETARIA DE VIGILÂNCIA SANITÁRIA. **Diretrizes de proteção radiológica em radiodiagnóstico médico e odontológico.** Portaria n.º 453 de 01 de junho de 1998. Brasília, DO de 02 de junho de 1998.

MOAYEDI, F.; BOOSTANI, R.; AZIMIFAR, Z.; SERAJEDIN KATEBI. **A Support Vector Based Fuzzy Neural Network approach for Mass Classification in Mammography.** In: 15th International Conference on Digital Signal Processing, p. 240-243, 2007.

NUNES, A. P.; SILVA, A. C.; PAIVA, A. C. **Detection of masses in mammographic images using geometry, Simpson's Diversity Index and SVM**. International Journal of Signal and Imaging Systems Engineering, v. 3, n. 1, p. 40-51, 2010.

OLSHAUSEN, B. A.; FIELD, D. J. **Emergence of simple-cell receptive-field properties by learning a sparse code for natural images.** Nature, v. 381, nº 13, p. 607-609, 1996b.

OLSHAUSEN, B. A.; FIELD, D. J. **Natural image statistics and efficient coding**. Network: Computation in Neural System, v. 7, p. 333-339, 1996a.

OLSHAUSEN, B.; FIELD, D. **What is the other 85% of V1 doing?** In: Problems in Systems Neuroscience. T. J. Sejnowski, L. van Hemmen, eds. Oxford University Press, 2004.

OTSU, N. **A Threshold Selection Method from Grey Level Histograms**, IEEE Transactions on Systems, Man and Cybernetics, v. 9, n. 1, p. 62-66, 1979.

OZEKES, S.; OSMAN, O.; ÇAMURCU, A. Y. **Mammographic Mass Detection Using a Mass Template**. Korean Journal of Radiology, v. 6, p. 221-228, 2005.

PANZERI, S; SCHULTZ, S. R.; TREVES, A; ROLLS, E. **Correlations and the encoding of information in the nervous system.** Proceeding Royal Society B: Biological Science. London: v. 266, nº 1423, p. 1001-1012, 1999.

PAPOULIS, A. **Probability, Random Variables and Stochastic Processes**. 4. ed. New York: McGraw-Hill, 2002.

PARKER, A. J.; HAWKEN, M. J. **Two-dimensional spatial structure of receptive fields in monkey striate cortex**. Journal of the Optical Society of America A, v. 5, nº 4, p. 598–605, 1988.

PENG-LU,YONGQIANG-LI,YUHE-TANG,ERYAN-CHEN. **Image Fault Area Detection Algorithm Based on Visual Perception.** International Journal of Image, Graphics and Signal Processing (IJIGSP), v. 3, n. 1, p. 24-30, 2011.

PRATT, W. K. **Digital Image Processing**. 3. ed. New York: John Wiley and Sons, 2001.

RENCHER, A. C. **Methods of multivariate analysis**. 2. ed. New York: Wiley-Interscience, 2002.

RINGACH, D. **Spatial Structure and Symmetry of Simple-Cell Receptive Fields in Macaque Primary Visual Cortex.** Journal of Neurophysiologic, v. 88, p. 455-463, 2002.

SAMPAIO, W. B.; DINIZ, E. M.; SILVA, A. C.; PAIVA, A. C.; GATTASS, M. **Detection of masses in mammogram images using CNN, geostatistic functions and SVM.** Computers in Biology and Medicine, v. 41, p. 653-664.

SCHÖLKOPF, B.; SMOLA, A.; MÜLLER, K. R. **Kernel principal component analysis**. Lecture Notes in Computer Science, v. 1327, p. 583-588, 1997.

SEIDE, F.; MERTINS, A. **Non-linear regression based feature extraction for connected-word recognition in noise.** In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'94), V. 2, P. 85-88, 1994.

SILVA, C. O. S. **Implementação de um sistema de telediagnóstico para classificação de massas em imagens mamográficas usando análise de componentes independentes.** 2012. 68p. Dissertação (Mestrado em Engenharia Elétrica / Automação e Controle), Universidade Federal do Maranhão – UFMA, São Luís.

SIMONCELLI, E. P. **Vision and the Statistics of the Visual Environment.** Current Opinion in Neurobiology, v. 13, p. 144-149, 2003.

SIMONCELLI, P; OLSHAUSEN, B. **Natural image statistics and neural representation.** Annual Review of Neuroscience, v. 24, p. 1193-1216, 2001.

SMITH, E. C.; LEWICKI, M. S. **Efficient auditory coding.** Nature, v. 439, n. 23, p. 978-982, 2006.

SOHNS, C.; ANGIC, B.; SOSSALLA, S.; KONIETSCHKE, F.; OBENAUER, S. **Computer-assisted Diagnosis in Full-field Digital Mammography—Results in Dependence of Readers Experiences**. The Breast Journal, v. 16, nº 5, p. 490–497, 2010.

SOUSA, C. M.; CAVALCANTE, A. B.; GUILHO, D.; BARROS, A. K. **Image Compression by Redundancy Reduction**. In: Proceedings of Independent Component Analysis and Blind Signal Separation, 7th International Conference (ICA 2007), Lecture Notes in Computer Science (LNCS), v. 4666, p. 422-429, London, 2007.

SUCKLING, J. et al.. **The Mammographic Image Analysis Society Digital Mammogram Database**. In: Exerpta Medica International Congress Series v. 1069 p. 375-378, 1994.

TCHOU, P. M. et al. **Interpretation Time of Computer-aided Detection at Screening Mammography**. Radiology, v. 257, nº 1, p. 40-46, 2010.

THULER, L. C. **Considerações sobre a prevenção do câncer de mama feminino.** Revista Brasileira de Cancerologia, v. 49, nº 4, p. 227-238, 2003.

TIMP, S.; VARELA, C; KARSSEMEIJER, N; **Temporal Change Analysis for Characterization of Mass Lesion in Mammography.** In: IEEE Transactions on Medical Imaging, v. 26, n. 7, p. 945-953, 2007.

WANG, Y.; GAO, X.; LI, J. **A Feature Analysis Approach to Mass Detection in Mammography Based on RF-SVM**. In: IEEE International Conference on Image Processing (ICIP 2007), v. 5, p. 9-12, 2007.

WEI, D.; CHAN, H. P.; HELVIE, M. A; SAHINER, B.; PETRICK, N.; ADLER, D. D.; GOODSITT, M. M. **Classification of Mass and Normal Breast Tissue on Digital Mammograms: Multiresolution Texture Analysis.** Medical Physics v. 22, nº 9, p. 1501-1514, 1995.

WILSON, A. D.; BOBICK, A. F. **Nonlinear Parametric Hidden Markov Models.** IEEE Journal on Robotics and Automation, Technical Report nº 424, 1997.

XU, S.; PEI, C. **Hierarchical Matching for Automatic Detection of Masses in Mammograms**. In: International Conference on Electrical and Control Engineering (ICECE), p. 4523-4526, 2011.

YUAN, Y.; GIGER, M. L.; LI, H.; SUZUKI, K.; SENNETT, C. **A dual-stage method for lesion segmentation on digital mammograms.** Medical Physics, v. 34, nº 11, p. 4180-4194, 2007.

ZHANG, G.; YAN, P.; ZHAO, H.; ZHANG, X. **A Computer Aided Diagnosis System in Mammography Using Artificial Neural Networks.** In: International Conference on Biomedical Engineering and Informatics. IEEE Computer Society, v. 2, p. 823-826, 2008.

ZHANG, L.; SANKAR, R.; QIAN, W. **Advances in microcalcification clusters detection in mammography**, Computer in Biology & Medicine, v. 32, nº 6, p. 515-528, 2002.

ZHANG, P.; VERMA, B.; KUMAR, K. **A Neural Genetic algorithm for feature selection and breast abnormality classification in digital mammography**. In: IEEE International Joint Conference on Neural Networks, v. 3, p. 2303-2308, 2004.

ZOHARIAN, S. A.; ROTHENBERG, M. **Principal-components analysis for low-redundancy encoding of speech spectra**. The Journal of the Acoustical Society of America, v. 69, nº 3, p. 832-845, 1981.

## GLOSSÁRIO

**Biópsia:** Operação que consiste em retirar um fragmento de tecido vivo de um órgão ou parte do corpo para exame histológico.

**Carcinoma:** Tumor canceroso.

**Ductos:** Canal; meio.

**Etiologia:** Parte da medicina que trata da origem das doenças.

**Hiperplasia:** Desenvolvimento anormal exagerado de um elemento anatômico ou de um tecido do organismo.

**Histograma:** Gráfico formado por retângulos de bases iguais, que correspondem a iguais intervalos de variável independente, e cujas alturas são proporcionais aos valores da grandeza em representação.

**Linfonodos:** Pequenos órgãos perfurados por canais que existem em diversos pontos da rede linfática, uma rede de ductos que faz parte do sistema linfático. Atuam na defesa do organismo humano e produzem anticorpos.

**Lóbulos:** Parte arredondada, saliente e pequena de um órgão.

**Lumpectomia:** É uma resposta cirúrgica comum para o câncer de mama. Envolve a remoção do tumor e de uma pequena quantidade de tecido saudável próximo à massa para ajudar a garantir a remoção de todos os vestígios do câncer.

**Mastectomia:** É uma resposta cirúrgica comum para o câncer de mama. Envolve a remoção do tumor e de todo o tecido saudável da mama.

**Menarca:** Primeiro período de menstruação.

**Metástase:** Mudança de forma ou de sede de uma afecção.

**Molibdênio:** Metal branco, elemento químico de símbolo *Mo*, número atômico 42 e massa atômica 95,95.

**Neoplasia:** Formação de um tecido novo de origem patológica; tumor.

**Nuliparidade:** Estado ou qualidade de nulípara. Fêmea que nunca pariu.

**Quimioterapia:** Tratamento de doenças por meio de substâncias químicas.

**Radioterapia:** Tratamento terapêutico pelos raios X ou por meio do rádio.

**Tecidos fibroglandulares:** Parênquima. Tecido conjuntivo que, nos animais, preenche os espaços que ficam entre outros tecidos ou certas cavidades do organismo.

**Tumores:** Formação patológica, não inflamatória, de tecido novo, que pode ser constituído por células normais e manter-se localizado (tumor benigno), ou ser formado por células atípicas, invadindo os tecidos vizinhos ou disseminando-se à distância (tumor maligno).