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

Luis Claudio de Oliveira Silva

Método de Detecção de Massas em Mamas Densas usando Análise de Componentes Independentes

São Luís
2017

Luis Claudio de Oliveira Silva

Método de Detecção de Massas em Mamas Densas usando Análise de Componentes Independentes

Tese apresentada ao Programa de Pós-graduação em Engenharia de Eletricidade da UFMA, como parte dos requisitos necessários para obtenção do grau de DOUTOR em Engenharia de Eletricidade.

Orientador: Prof. Dr. Allan Kardec Duailibe Barros Filho

Universidade Federal do Maranhão

São Luís

2017

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Núcleo Integrado de Bibliotecas/UFMA

Silva, Luis Claudio de Oliveira.
Método de Detecção de Massas em Mamas Densas usando Análise de Componentes Independentes / Luis Claudio de Oliveira Silva. – 2017.
70 f.

Orientador(a): Allan Kardec Duailibe Barros.
Tese (Doutorado) – Programa de Pós-graduação em Engenharia de Eletricidade/ccet, Universidade Federal do Maranhão, Laboratório PIB 3/CCET/UFMA, 2017.

1. Agrupamento. 2. Análise de Imagem Médica. 3. Filtragem. 4. Mamas Densas. 5. Segmentação de Imagens. I. Barros, Allan Kardec Duailibe. II. Título.

Luis Claudio de Oliveira Silva

Método de Detecção de Massas em Mamas Densas usando Análise de Componentes Independentes

Tese apresentada ao Programa de Pós-graduação em Engenharia de Eletricidade da UFMA, como parte dos requisitos necessários para obtenção do grau de DOUTOR em Engenharia de Eletricidade.

Aprovado em 27 de julho de 2017

BANCA EXAMINADORA

Prof. Dr. Allan Kardec Duailibe Barros Filho

Universidade Federal do Maranhão

Prof. Dr. Ewaldo Eder Carvalho Santana

Universidade Federal do Maranhão

Prof. Dr. João Viana da Fonseca Neto

Universidade Federal do Maranhão

Profa. Dra. Áurea Celeste da Costa Ribeiro

Universidade Estadual do Maranhão

Prof. Dr. Fausto Lucena de Oliveira

Universidade Ceuma

Ao Lucas, Henrique e Luísa.

Agradecimentos

Ao meu orientador Prof. Allan Kardec, pela confiança e pela oportunidade concedida.

Ao Prof. Ewaldo Santana, por toda a confiança creditada e pelos ensinamentos sobre objetividade.

Ao Programa de Pós-Graduação em Engenharia de Eletricidade da UFMA.

Aos colegas do Laboratório de Processamento da Informação Biológica (PIB), em especial a Marcus Vinicius, que esteve comigo nessa caminhada sempre compartilhando conhecimento.

Aos colegas do curso de Engenharia da Computação da UFMA.

À CAPES pelo financiamento deste trabalho.

Resumo

SILVA, L. C. de O. **Método de Detecção de Massas em Mamas Densas usando Análise de Componentes Independentes**. 2017. 70 f. Tese de Doutorado - Programa de Pós-Graduação em Engenharia de Eletricidade, Universidade Federal do Maranhão, São Luís, 2017.

O câncer de mama é o segundo tipo de câncer que mais afeta mulheres no mundo, perdendo apenas para o câncer de pele não melanoma. A densidade da mama pode dificultar a localização de massas, especialmente em estágios iniciais. Neste trabalho, propõe-se o uso de análise de componentes independentes para detectar e segmentar lesões em mamas densas. Vários trabalhos sugerem o uso do diagnóstico auxiliado por computador, aumentando a sensibilidade para acima de 90% na detecção de câncer em mamas não densas, no entanto, existem poucos estudos publicados sobre a detecção em mamas densas. Para analisar a eficiência do método proposto em relação a outras técnicas de segmentação, comparamos o desempenho com a análise de componentes principais. Para medir a qualidade da segmentação obtida pelos dois métodos, será utilizada uma medida de sobreposição de área. Para verificar se houve diferença entre os resultados dos métodos propostos na detecção de lesões em mamas não densas e nas mamas densas, foi utilizado um teste estatístico para duas proporções. Os resultados experimentais usando os bancos de dados Mini-MIAS e DDSM mostraram uma acurácia de 92,71% na detecção de massas em mamas não densas e 79,17% em mamas densas. Todas as experiências mostraram que os filtros de ICA usados têm um melhor desempenho para detectar lesões em mamas densas, em comparação com PCA. Contrariamente aos trabalhos anteriores, nossos experimentos mostraram que existe realmente uma diferença significativa entre a detecção de massas em mamas densas e não densas. Este estudo pode ajudar o especialista a detectar lesões em mamas densas.

Palavras-chave: Análise de Imagem Médica, Mamas Densas, Filtragem, Agrupamento, Segmentação de Imagens.

Abstract

SILVA, L. C. de O. **Method for Detection Masses in Dense Breast using Independent Component Analysis**. 2017. 70 p. Doctoral Thesis - Postgraduate Program in Electrical Engineering, Federal University of Maranhão, São Luís, 2017.

Breast cancer is the second type of cancer that most affects women in the world, losing only for non melanoma skin cancer. Breast density can hinder the location of masses, especially in early stages. In this work, the use of independent component analysis for detecting and segmentation lesions in dense breasts is proposed. Several works suggests the use of computer aided diagnosis, increasing sensitivity to over 90% in detecting cancer in non dense breasts, however there are few published studies about detecting in dense breasts. To analyse its efficiency in relation to other segmentation techniques, we compare the performance with principal component analysis. To measure the quality of the segmentation obtained by the two methods, a area overlay measure will be used. To verify if there was any difference between the results of the proposed methods in the detection of lesions in nondense breasts and in dense breasts, a statistic test for two proportions was used. Experimental results on the Mini-MIAS and DDSM database showed an accuracy of 92.71% in detecting masses in nondense and 79.17% in dense breasts. All experiments showed that the ICA filters have a better performance for detect lesions in dense breast, compared with PCA. Contrary to previous works, our experiments showed that there is actually a significant difference between the detection of masses in dense and nondense breasts. This study can help specialist to detect lesions in dense breast.

Keywords: Medical Image Analysis, Dense Breast, Filtering, Clustering, Image Segmentation.

Sumário

Lista de Abreviaturas

ICA	Independent Component Analysis
PCA	Principal Component Analysis
CAD	Diagnóstico Auxiliado por Computador
ANN	Artificial Neural Networks
ROI	Regions of Interest
MIAS	Mammographic Image Analysis Society
DDSM	Digital Database for Screening Mammography
EICAMM	Enhanced ICA Mixture Model
SVM	Support Vector Machine
RBFNN	Radial Basis Function Neural Network
RNA	Redes Neurais Artificiais
SOM	Self Organizing Maps
BI-RADS	Breast Image Reporting and Data System
CC	Crânio Caudal
MLO	Médio Oblíquo Lateral
AOM	Area Overlap Measure
BSS	Blind Source Separation
DFT	Transformada Discreta de Fourier
DCT	Transformada Discreta do Cosseno
ACR	Colégio Americano de Radiologia

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O câncer de mama é um dos tipo de câncer que mais acomete mulheres no mundo. As estimativas para os casos de câncer no Brasil nos anos de 2016 e 2017 indicam aproximadamente 600 mil novos casos (INCA, 2007). Destes casos, os mais incidentes ainda são o câncer de pele do tipo não melanoma (58 mil novos casos), o câncer de próstata (61 mil novos casos) e o câncer de mama em mulheres (58 mil novos casos). O câncer de mama também acomete homens, porém é raro, representando apenas 1% do total de casos da doença.

A mamografia é um dos exames mais utilizados para detectar o câncer de mama e sua efetividade aumenta quando é realizada no estágio inicial da doença (GALLARDO-CABALLERO et al., 2012). Embora seja bastante utilizado e tenha sensibilidade variando de 46% a 88%, e a utilização desse exame como método de rastreamento reduza a mortalidade em 25%, a detecção de lesões em mamografias depende de fatores tais como tamanho e localização da lesão, densidade da mama, qualidade da mamografia e habilidade de interpretação do especialista (INCA, 2007).

A segmentação da imagem é o primeiro passo no processamento de imagens médicas (RIBEIRO et al., 2013). No caso das imagens de mamografia, as lesões são segmentadas e avaliadas por um classificador. Assim, é importante que a segmentação possa detectar lesões nas situações mais difíceis, como é o caso de mamas de maior densidade.

Vários métodos, baseados em sistemas de diagnóstico auxiliado por computador (CAD) têm sido propostos para auxiliar os especialistas na identificação de lesões, fornecendo uma segunda opinião no processo de identificação. Estes sistemas empregam técnicas de processamento de imagens e reconhecimento de padrões, capazes de produzir resultados mais confiáveis, podendo detectar lesões e microcalcificações logo no início, aumentando dessa forma a chance de cura.

1.1 Trabalhos Relacionados

Christoyianni et al. (2002) desenvolveram um sistema CAD baseado em análise de componentes independentes (ICA) associada com redes neurais artificiais (ANN), para classificar regiões normais ou suspeitas, obtendo acurácia de 88,23%. Abu-Amara e Abdel-Qader (2007) propuseram um método para a detecção de regiões suspeitas (ROS) baseado em extração de características extraídas com ICA, obtendo acurácia de 79,00% na detecção de anomalias e 71,20% no diagnóstico de massa. Abdel-Qader e Abu-Amara (2008) apresentaram um método baseado em lógica nebuloza para identificar regiões suspeitas, obtendo acurácia de 84,03%.

Campos et al. (2011) apresentaram um método baseado em ICA para detectar regiões suspeitas, obtendo 90,15% de acerto na classificação entre tecidos normais e suspeitos. Gallardo-Caballero et al. (2012) desenvolveram um método para detectar microcalcificações em mamas usando ICA com 91,8% de sensibilidade, usando um banco de dados digital para rastreamento de mamografias (DDSM) (HEATH et al., 2001). Ribeiro et al. (2013) propuseram a segmentação automática de massas usando uma técnica denominada Enhanced ICA Mixture Model (EICAMM) com 3,54% de taxa de erro para detectar massas.

García-Manso et al. (2013) analisaram se a densidade do tecido mamário afeta a detecção de massas em mamografias. Em seu trabalho, usou ICA para extrair características, redes neurais artificiais e máquina de vetor de suporte (SVM) para detectar massas em regiões de interesse (ROIs) extraídas de imagens de mamas de diferentes densidades, obtendo acurácia de 88,41%. Seus melhores resultados mostraram uma área sob a curva ROC de 0,965 para classificação de casos de mama de baixa densidade e 0,897 para casos de mama de maior densidade e sua conclusão é que o desempenho de seu método é afetado pela densidade mamária. A Tabela 1.1 sumariza os trabalhos citados.

Tabela 1.1: *Trabalhos relacionados*

Trabalho	Técnica	Classificador	Base de dados	ROIs	Acerto
Christoyianni	ICA	RBFNN	MIAS	238	88.23%
Abu-Amara	ICA	SVM	MIAS	119	79%
Abdel-Qader	ICA	Fuzzy	MIAS	119	84,03%
Campos	ICA	RNA	MIAS	150	90.15%
Gallardo-Caballero	ICA	ANN	DDSM	100	91.8%
Ribeiro	EICAMM	SOM	DDSM/LAPIMO	396	46,71%
Garcia-Manso	ICA	SVM	DDSM	5052	88,41%

De acordo com Brem. et al. (2005), a detecção de câncer de mama não é impactada pela densidade mamária. Em seus experimentos, 89% dos casos de câncer foram detectados por um sistema CAD; 90% dos casos de câncer em mamas não densas e 88% dos casos de câncer em mamas densas. Oliver et al. (2010) estudaram o desempenho de um sistema CAD para detecção de câncer de mama. De acordo com seus estudos, quando a informação da densidade mamária não foi considerada, a sensibilidade obtida pelo sistema CAD foi de 74,7% para as visualizações do tipo crânio caudal (CC) e 85,3% para as visualizações do tipo médio oblíquo lateral (MLO). Considerando a informação da densidade mamária, a sensibilidade para mamografias do tipo CC e MLO aumenta para 80,0% e 89,3%, respectivamente, e 82,7% e 90,7%, respectivamente, usando estimação automática.

Assim, de acordo com Brem. et al. (2005), não houve diferenças estatisticamente significativas na detecção de massas em mamas densas ou não densas. No entanto, de acordo com Oliver et al. (2010), a informação da densidade mamária pode melhorar os sistemas CAD na tarefa de detecção de massa. Apesar de haverem muitos trabalhos publicados sobre ferramentas desenvolvidas para localização e classificação de lesões, poucos são direcionados para detecção em mamas consideradas densas.

Existem muitos trabalhos publicados sobre ferramentas desenvolvidas para localização e classificação de massas usando ICA. De fato, ICA surgiu como uma técnica que tem uma inspiração biológica muito forte (OLSHAUSEN; FIELD, 1996; SMITH; LEWICKI, 2006) com muitas aplicações (HYVARINEN, 1999; HYVARINEN; OJA, 2000; HYVARINEN; KARHUNEN; OJA, 2004; JAMES; HESSE, 2005; MOK; LAM; NG, 2004; CHIEN; CHEN, 2006; HOYER; HYVARINEN, 2000; OLSHAUSEN; FIELD, 1996; BARTLETT; MOVELLAN; SEJNOWSKI, 2002; CAVALCANTE et al., 2009; JENSSEN; ELTOFT, 2003). Neste trabalho, foi avaliado se a densidade da mama afeta o resultado da detecção de massas em mamas densas. Para isso, foi usado codificação eficiente baseada em ICA para extração de características e detecção de massas em mamas tanto tanto em mamas não densas como em mamas de alta densidade. Para esta tarefa, foram obtidas ROIs dos bancos de dados Mini-MIAS (SUCKLING et al., 1994) e DDSM (HEATH et al., 2001), onde as características que definem classes de imagens contendo massas foram extraídas. Essas características foram então usadas para gerar um banco de filtros usado para segmentar massas em mamas densas.

Para verificar a eficiência do método proposto em relação a outras técnicas de segmentação, foi efetuada uma comparação do desempenho com PCA. Para medir a qualidade da segmentação obtida pelos dois métodos, a medida de sobreposição de área (*Area Overlay Measure*, AOM) foi usado. Para verificar se houve alguma diferença entre os resultados dos métodos na detecção de lesões em mamas não densas e nas mamas densas, utilizou-se o teste Z de hipóteses para duas proporções.

1.2 Objetivos

O objetivo geral deste trabalho é desenvolver um método para identificação e classificação de massas em imagens de mamografia de mamas densas, a partir de técnicas de processamento de imagem e reconhecimento de padrões. Para alcançar o objetivo geral pretendido, buscar-se-á atingir os seguintes objetivos específicos:

- Avaliar a técnica de análise de componentes independentes como descritor de anormalidades do tecido mamário;
- Desenvolver um método para extração e seleção de características de massas a partir de imagens de mamografia;
- Segmentar regiões de interesse em mamografias utilizando técnicas de extração de características bem como um banco de filtros, auxiliado por um algoritmo de agrupamento; a partir da imagem de mama densa, previamente classificada, propor um algoritmo para efetuar a localização de massas em imagens mamográficas de mamas densas;
- Comparar a efetividade dos algoritmo com outras técnicas de abordadem semelhante (análise de componentes principais)
- Verificar se há diferença significativa entre a detecção em imagens de mamas densas e imagens de mamas não densas;
- Implementar os algoritmos propostos e desenvolver um sistema capaz de auxiliar especialistas da área médica/radiológica no trabalho de localização e segmentação de achados em imagens mamográficas de mamas densas.

1.3 Contribuições

Este trabalho apresenta as seguintes contribuições:

- Avalia a tarefa de segmentação de massas em imagens de mamas densas a partir de filtros de ICA e PCA;
- Apresenta um método para detecção e segmentação de massas em imagens de mamografia de mamas densas para auxílio ao diagnóstico do especialista;
- Avalia se há diferença significativa entre a detecção de massas em mamas densas e em mamas não densas usando filtros de ICA.

1.4 Organização do trabalho

Este trabalho é organizado da seguinte forma. O Capítulo 2 trata da fundamentação teórica necessária para detecção de nódulos em mamografias de mamas densas. Serão abordados os conceitos referentes a análise de imagens mamográficas, sistemas computacionais para auxílio ao diagnóstico, processamento digital de imagens, extração de características de imagens usando ICA e PCA e a técnica de agrupamento aplicada na segmentação da região de interesse na mamografia.

O capítulo 3 descreve a metodologia empregada para efetuar a aquisição das imagens usadas nos experimentos, o treinamento do algoritmo de detecção e os testes realizados.

O capítulo 4 aborda os resultados obtidos nos experimentos realizados bem como faz algumas discussões acerta dos resultados e análises estatísticas pertinentes. O capítulo 5 efetua uma discussão dos resultados obtidos nos experimentos.

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

Capítulo 2

Fundamentação Teórica

2.1 Câncer de mama

O câncer de mama pode ser definido como sendo uma disfunção nas células que compõem o tecido mamário, capaz de ocasionar multiplicação celular desordenada e causando o surgimento de estruturas benignas ou malignas. As estruturas malignas podem se espalhar para outras regiões do corpo, invadindo outros órgãos. O termo neoplasia é utilizado para alterações celulares que acarretam no crescimento desordenado destas células, podendo ser benigna ou maligna (INCA, 2007).

As neoplasias malignas se dividem, por sua vez, em tumores epiteliais, ou carcinomas, de origem ductal ou lobular, e em sarcomas, originados a partir de tecidos conjuntivos. Os carcinomas ductais formam-se nos ductos que levam o leite do lóbulo para os mamilos. Já os carcinomas lobulares formam-se nos bulbos que produzem o leite materno. A Figura 2.1,(a) e (b), ilustra a formação de carcinomas lobulares e ductais (ACS, 2017). Os carcinomas lobulares acometem os lóbulos, glândulas produtoras de leite no final dos ductos mamários. Os carcinomas ductais ocorrem nos ductos mamários, canais por onde passa o leite até os mamilos.

Assim como qualquer tipo de câncer, o câncer de mama pode se espalhar para outras partes do corpo, processo este conhecido como metástase. Por isso é de vital importância que sua detecção seja a mais precoce possível, aumentando assim a chance de tratamento e de cura. As formas mais eficazes para detecção precoce do câncer de mama são o exame clínico e a mamografia.

A manifestação do câncer pode se dar através do surgimento de massas e calcificações. As massas, ou nódulos, correspondem a um conjunto de células aglomeradas, tornando-se mais densas do que o tecido ao redor, podendo configurar neoplasias benig-

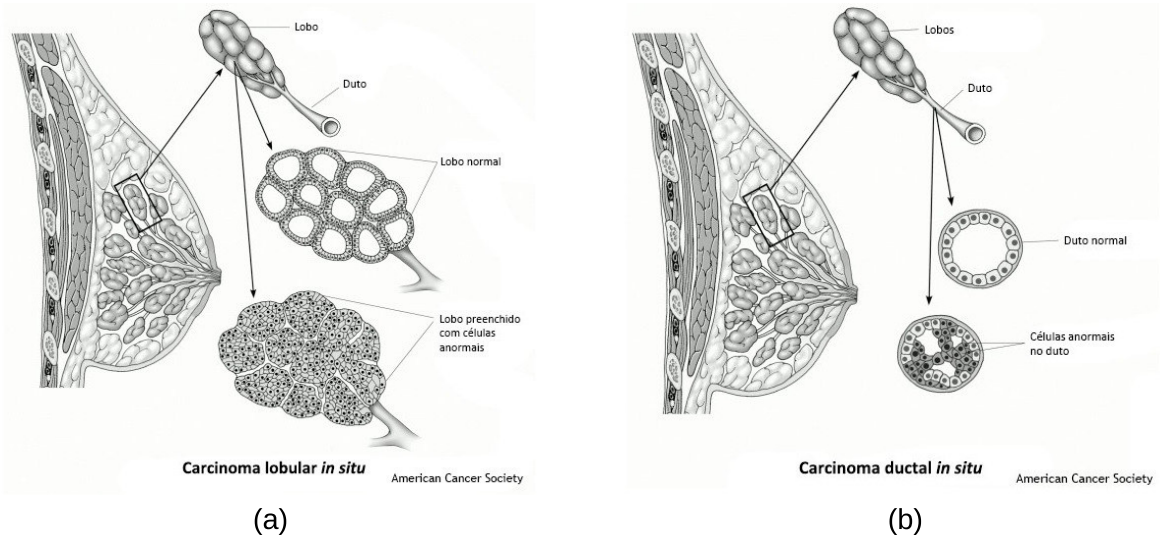

Figura 2.1: *Formação de carcinomas lobulares e ductais na mama. Fonte: ACS (2017)*

nas ou malignas. As calcificações correspondem a depósitos de sais de cálcio desenvolvidos no tecido mamário, que podem inclusive ser consequências de processos inflamatórios, alterações degenerativas, processos tóxicos metabólicos, traumatismos ou resultados de processos secretores ativos de células tumorais.

Na Figura 2.2, pode-se visualizar duas imagens de neoplasias, extraídas de mamografias de dois pacientes. Na Figura 2.2(a), uma neoplasia benigna; na Figura 2.2(b), uma neoplasia maligna. As imagens foram obtidas da base de dados Mini-MIAS.

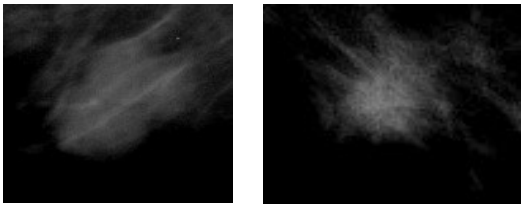

Figura 2.2: *Exemplos de neoplasias: (a) benigna; (b) maligna. Fonte: Suckling et al. (1994)*

Na Figura 2.3, pode-se visualizar um exemplo de neoplasia do tipo calcificação: (a) região de interesse, sem realce para melhorar a visualização; (b) mesma região de interesse, com realce para melhorar a visualização das calcificações.

2.2 Mamografia

A mamografia é uma forma de radiografia da mama, destinada a registrar imagens da mama com a finalidade de diagnosticar a presença de anomalias na mama. A mamografia é realizada por um equipamento denominado mamógrafo, que comprime as mamas

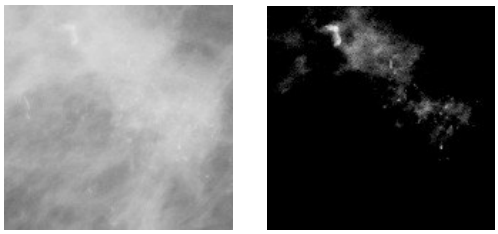

Figura 2.3: *Exemplo de neoplasia do tipo calcificação: (a) imagem original; (b) imagem realçada. Fonte: Suckling et al. (1994)*

com o objetivo de permitir uma melhor visualização de pequenas alterações, muitas vezes difíceis de serem detectadas a olho nu (INCA, 2007). A detecção nos estágios iniciais aumenta a chance de tratamento e a radiografia da mama permite detectar alterações ainda neste estágio. Na Figura 2.4 é possível visualizar um mamógrafo para realização de exames de mamografia.

Figura 2.4: *Mamógrafo. Na mamografia, cada mama é comprimida horizontalmente e, em seguida, obliquamente, armazenando uma imagem de raios-X de cada posição. Fonte: INCA (2007)*

Existem dois tipos de mamografia: a convencional e a digital. Apesar dos dois sistemas utilizarem raios-X para produzir a imagem da mama, a diferença está em como a imagem é captada. A mamografia convencional utiliza um filme para a exposição da imagem após a exposição da mama aos raios-X. Já a mamografia digital possui um sensor que transforma os raios-X em sinais elétricos, transmitindo-os para um computador,

possibilitando seu armazenamento e processamento a partir de programas específicos de análise de imagens mamográficas.

A identificação de estruturas que possam indicar a presença de anomalias se dá através da constatação de uma diferença de contraste entre os diversos tecidos envolvidos. A gordura, por exemplo, absorve uma menor quantidade de raios-X, aparecendo mais escura no mamograma, enquanto tecidos fibroglandulares apresentam densidade óptica maior e aparecem mais claros (BOYD et al., 2007). Geralmente, microcalcificações e massas aparecem em tonalidades mais claras na imagem obtida após a revelação do filme mamográfico, 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 (INCA, 1994).

Para cada paciente, duas imagens de cada mama são submetidas à este exame. Cada incidência de raios-X irá gerar uma visão, denominadas crânio caudal (CC) e médio oblíquo lateral (MLO). Na Figura 2.5, é apresentada uma ilustração da realização deste exame e a obtenção de cada uma das duas visões de uma das mamas.

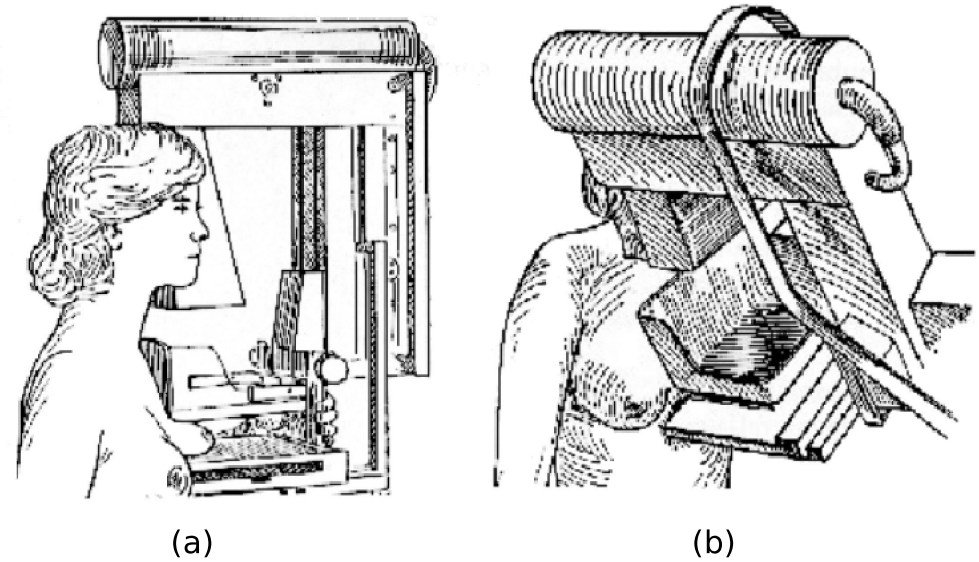

Figura 2.5: *Ilustração da realização de uma mamografia e obtenção das visões crânio caudal (CC) e médio oblíquo lateral (MLO) de uma das mamas. Fonte: Angelo (2007).*

Pode-se observar nas Figuras 2.6 (a)-(d) exemplos de mamografias nas visões crânio caudal (CC) e médio oblíquo lateral (MLO) de um único paciente.

A composição do tecido mamário pode ser um complicador para na detecção de lesões em mamografias. O tecido adiposo possui menor densidade (razão entre o tecido

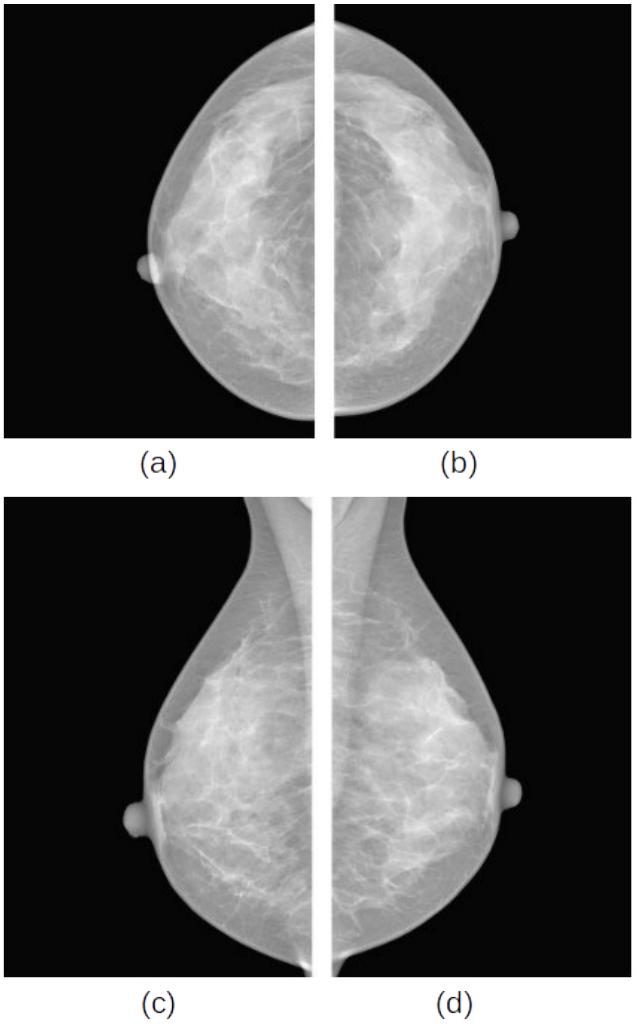

Figura 2.6: *Exemplos de mamografia: (a) visão crânio caudal (CC) da mama direita; (b) visão CC da mama esquerda; (c) médio oblíquo lateral (MLO) da mama direita; (d) visão MLO da mama esquerda. Fonte: Moreira et al. (2012)*

fibroglandular e o tecido adiposo da mama) e, consequentemente, permite melhor detecção das lesões. Já o tecido fibroglandular é mais denso e, por essa razão, é mais difícil detectar as lesões contidas neste tipo de tecido.

Vários fatores podem influenciar na modificação da composição do tecido da mama. Tais mudanças podem ser causadas por alterações hormonais ou predisposição genética. Mulheres jovens apresentam composição mamária feita predominantemente de tecido glandular e pouca gordura. No entano, é possível ocorrer de mulheres mais velhas apresentarem mamas extremamente densas. Além disso, ganho ou perda de peso também podem influenciar na composição da mama e consequentemente, afetar sua densidade (GARCÍA-MANSO et al., 2013).

Foram propostos métodos para avaliação da densidade mamográfica, e dos métodos existentes, um dos mais utilizados é a classificação proposta pelo Colégio Americano de

Radiologia (ACR) (BERG et al., 2000), denominada Breast Imaging Reporting and Data System (BI-RADS) (ACR, 1998), ou escala BI-RADS. O ACR desenvolveu um conjunto de recomendações para a padronização de relatórios de mamografia, ultra-sonografia (USG) e ressonância magnética (MR). A escala BI-RADS permite a padronização de relatórios de mamografia, sujeita a confusão na interpretação de resultados quando se utilizam critérios puramente descritivos (ACR, 1998). Além de descrever os achados radiológicos, a escala BI-RADS classifica a composição mamária em quatro categorias. Esta classificação é utilizada em laudos mamográficos, facilitando a identificação por parte dos especialistas e aplicação em estudos epidemiológicos.

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

- BI-RADS I: Mamas adiposas, contendo cerca de até 25% do componente fibroglandular;
- BI-RADS II: Mamas predominantemente adiposas, contendo cerca de 26% até 50% do componente fibroglandular;
- BI-RADS III: Mamas com padrão denso e heterogêneo, nas quais se observa 51 a 75% de tecido fibroglandular, o que pode dificultar a visualização de eventuais nódulos;
- BI-RADS IV: Mamas muito densas, por apresentarem mais de 75% de tecido fibroglandular, o que pode diminuir a sensibilidade da mamografia.

Na Figura 2.7, pode-se ver quatro mamografias, classificados segundo o sistema BI-RADS.

2.3 Diagnóstico auxiliado por computador

Com o objetivo de fornecer ao especialista na área médica ferramentas para auxiliar na detecção e diagnóstico de lesões em imagens mamográficas, diversos trabalhos propõem o desenvolvimento de sistemas computacionais automáticos ou semiautomáticos, capazes de fornecer uma sugestão de diagnóstico. Estas ferramentas são denominadas de Sistemas de Auxílio à Detecção/Diagnóstico ou simplesmente Sistemas CAD/CADx (Computer-Aided Detection/Diagnosis) e são capazes de fornecer uma segunda opinião para o especialista, reduzindo erros de interpretação causados por má qualidade da imagem ou grande volume de exames a serem analisados para poucos profissionais disponíveis.

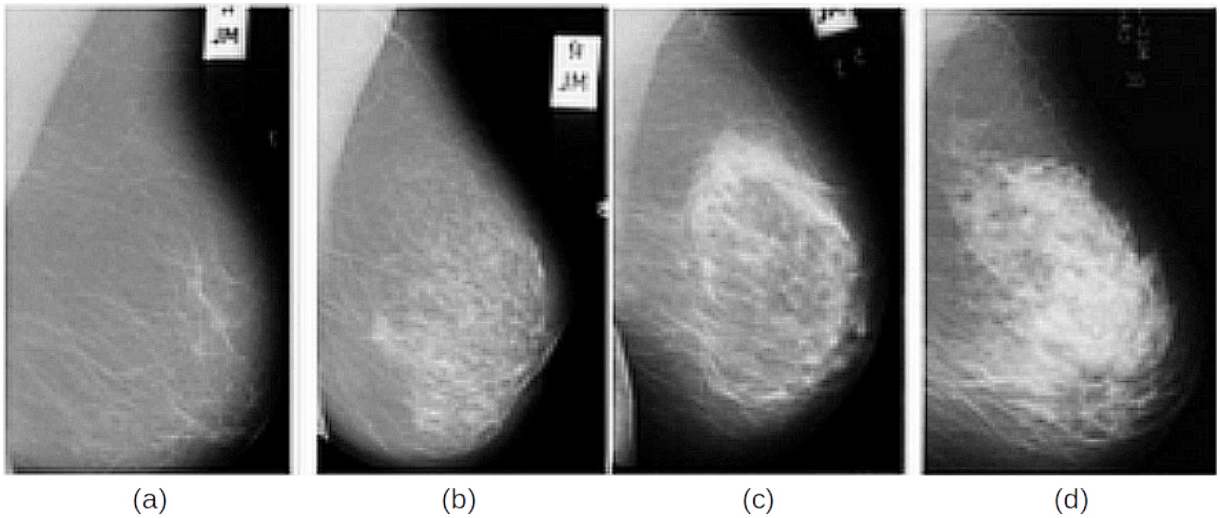

Figura 2.7: *Exemplos de mamas classificadas quanto a densidade, de acordo com a escala BI-RADS. (a) BI-RADS I, (b) BI-RADS II, (c) BI-RADS III e (d) BI-RADS IV. Fonte: Silva e Menotti (2012)*

Estes trabalhos tiveram início em 1967, com a elaboração de um procedimento para analisar a densidade óptica de imagens mamográficas, e a partir desta análise identificar áreas suspeitas (WINSBERG et al., 1967). Porém, foi com o desenvolvimento de sistemas radiológicos mais modernos e técnicas mais avançadas de processamento de imagens, que os sistemas CAD passaram a produzir resultados mais confiáveis, podendo até detectar regiões de microcalcificações.

Os sistemas CAD/CADx são capazes de fornecer uma segunda opinião ao especialista, algoritmos e computadores. Em se tratando de análise de imagens de mamografia, por exemplo, estes sistemas podem detectar regiões suspeitas em uma mamografia, além de analisar estas regiões quanto às suas características de benignidade e malignidade, auxiliando no apoio à decisão do diagnóstico médico (TANG et al., 2009).

Nos sistemas CAD/CADx baseados em tecnologia de mamografia em filme radiográfico, as mamografias são digitalizadas utilizando scanners de alta resolução óptica, algo em torno de 3200 pontos por polegada (dpi, do inglês dot per inch), e armazenados em computadores para análise. Já nos sistemas baseados em tecnologia de mamografia digital, a imagem é adquirida através de um receptor digital e enviada diretamente para o computador, ao invés de usar um filme radiográfico. Desta forma, a mamografia digital possibilita maior rapidez na conclusão do exame, bem como mais eficiência na detecção de lesões ao passo que diminui a taxa de ruído acrescentado na imagem enviada para o computador.

Após a aquisição, a imagem mamográfica é sujeita a técnicas de análise baseadas em processamento de imagem para remoção de artefatos e redução de ruídos. Na Figura 2.8,

é possível observar o resultado da remoção de artefatos em uma imagem de mamografia.

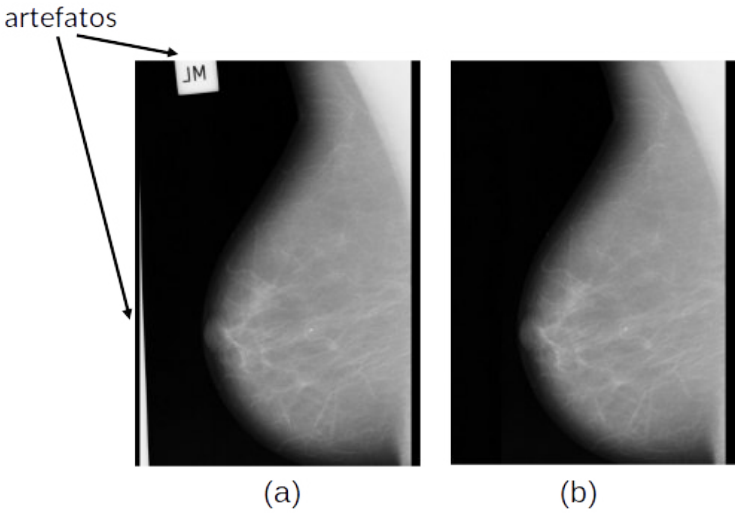

Figura 2.8: *Exemplo de remoção de artefatos em mamografia. (a) mamografia com artefatos; (b) mamografia sem artefatos. Fonte: Suckling et al. (1994)*

Uma vez que estes artefatos seja removidos, comumente são aplicados métodos de detecção de massas ou lesões nas imagens, com o objetivo de encontrar regiões suspeitas que possam ser, posteriormente, classificadas em benignas ou malignas.

Em se tratando de mamas densas, existem poucos sistemas CAD específicos. Segundo Yang et al. (2007), o número de falso-positivos cresce com o aumento da densidade mamária. De acordo com Obenauer et al. (2006), a densidade mamária pode afetar a detecção por CAD. Ho e Lam (2003) mostraram redução da sensibilidade estatística do CAD com o aumento da densidade mamária. Em seus experimentos, Ho e Lam (2003) notaram que a sensibilidade do CAD foi de 93,3%, com especificidade de 1,3 falsos positivos por imagem nos casos de mamas adiposas. Entretanto, a sensibilidade para 64,3% (especificidade de 1,2) para mamas muito densas.

2.4 Processamento digital de imagens

O processamento digital de imagens compreende um conjunto de procedimentos executados em uma imagem por meio de hardware e software, bem como fundamentos teóricos usados na elaboração de algoritmos usados em tarefas como filtragem, reconhecimento de padrões, representação, dentre outras (GONZALEZ; WOODS; EDDINS, 2003).

Uma imagem digital pode ser representada matematicamente por uma função $f(x, y)$, onde x e y representam as coordenadas espaciais e o valor de f em qualquer ponto, o brilho ou nível de cinza da imagem naquele ponto. Desse modo, uma imagem digital é

uma imagem discretizada tanto em coordenadas espaciais quanto em brilho (GONZALEZ; WOODS; EDDINS, 2003), sendo comumente representada por uma matriz cujos índices de linha e coluna identificam um ponto na imagem, e o correspondente valor o brilho ou nível de cinza, naquele ponto. Os elementos desta matriz são comumente denominados pixels. A Figura 2.9 mostra a representação de uma imagem de tamanho M x N.

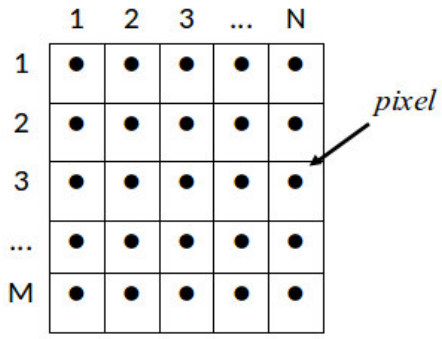

Figura 2.9: *Representação de uma imagem de tamanho M x N como matriz.*

Os passos fundamentais para efetuar o processamento de uma imagem encontram-se ilustrados na Figura 2.10. De acordo com o fluxograma apresentado na Figura, o primeiro passo é a aquisição da imagem, através de algum dispositivo capaz de capturar a imagem, transformando-a em uma imagem digital capaz de ser processada por um computador ou dispositivo com este fim. Este dispositivo pode ser, por exemplo, um scanner ou uma câmera digital.

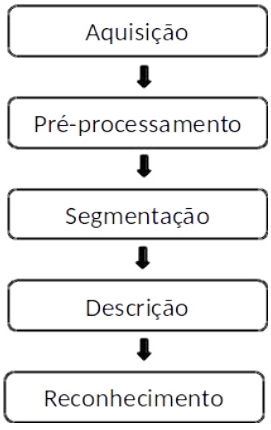

Figura 2.10: *Passos para o processamento de imagens.*

Logo após a aquisição, pode-se opcionalmente arquivar a imagem, em um dispositivo de memória temporário, tal qual a memória RAM do computador, ou em um dispositivo de memória permanente, como por exemplo, um disco rígido ou dispositivo de armazenamento baseado em memória flash.

Após, o próximo passo é o pré-processamento. Nesta fase, faz-se o melhoramento da imagem, aumentando a chance de sucesso nas etapas seguintes (GONZALEZ; WOODS;

EDDINS, 2003). Estão incluídas nesta etapa, tarefas como realce de contraste, remoção de imperfeições ou ruídos, dentre outras.

A segmentação divide a imagem de entrada em partes ou regiões, com o objetivo de simplificar a tarefa de processamento através da localização de objetos de interesse e formas na imagem. Como resultado, pode-se obter um conjunto de regiões ou um conjunto de contornos, com alguma característica ou propriedade de interesse para o usuário.

A descrição, também denominada seleção de características, extrai características que resultem em informação quantitativa de interesse ou que permitam a discriminação entre classes de objetos, tais como forma e textura. Dessa forma, pode-se efetuar o reconhecimento destes objetos, através da atribuição de um rótulo, ou classe, a cada objeto baseado na informação dada pelo descritor. Dessa forma pode-se, por exemplo, identificar caracteres em uma imagem, objetivando reconhecer letras e palavras que ali possam existir.

Estes procedimentos podem ser aplicados em imagens mamográficas para:

- efetuar filtragem, removendo ruídos ou informação desnecessária incorporado no processo de digitalização da mamografia;
- detecção de objetos, tais como calcificações, massas, distorções arquiteturais e assimetria bilateral;
- classificação de massas ou microcalcificações encontradas através de descritores de forma ou textura;
- permitir registro e recuperação de imagens de mamografia baseados em conteúdo, análogo à motores de busca textual.

2.5 Análise de componentes independentes

A análise de componentes independentes (ICA) é um método baseado em estatística computacional desenvolvido, inicialmente, para resolver problemas de separação cega de fontes (BSS) (HYVARINEN; OJA, 2000). Seu objetivo é separar um conjunto de sinais obtidos a partir de sinais misturados, sem para isso, conhecer alguma informação acerca dos sinais originais, ou fontes, e do processo de mistura (CHOI et al., 2005).

Trata-se de um modelo generativo, pois os sinais misturados são a combinação linear dos sinais originais (componentes independentes) com uma matriz de mistura, pos-

suindo aplicações em diversas áreas: áudio, sinais de SONAR, instrumentação médica, comunicação móvel, engenharia biomédica, dentre outras.

Neste modelo, a análise ou separação das fontes ou sinais originais que são estatisticamente independentes é efetuada a partir de um determinado modelo de mistura das fontes. Outra aplicação para ICA é a extração de características. Em processamento de imagem, as componentes podem fornecer uma representação para uma imagem. Tal representação permite executar tarefas como compressão ou reconhecimento de padrões.

Em ICA, os dados são dispostos na forma matricial $X = [x_1, x_2, ..., x_M]^T$, sendo $x_1, x_2, ..., x_M$ misturas de variáveis estatisticamente independentes desconhecidas $S = [s_1, s_2, ..., s_N]^T$. Tal mistura é obtida através de uma transformação linear. Na forma matricial, esta transformação é representado como

$$X = AS. \tag{2.1}$$

Na Figura 2.11, pode-se ver dois sinais artificiais estatisticamente independentes que, depois de misturados, produzem os sinais apresentados na Figura 2.12. No contexto de separação cega de fontes, estes sinais artificiais compõem as linhas da matriz S.

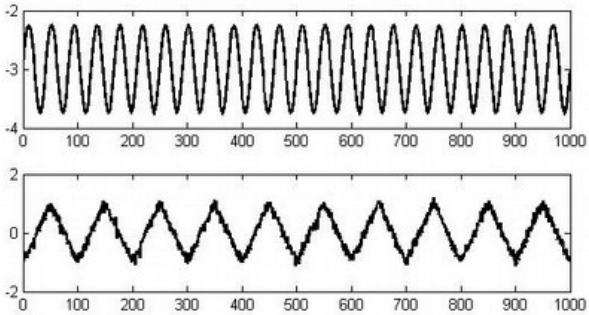

Figura 2.11: *Sinais artificiais gerados.*

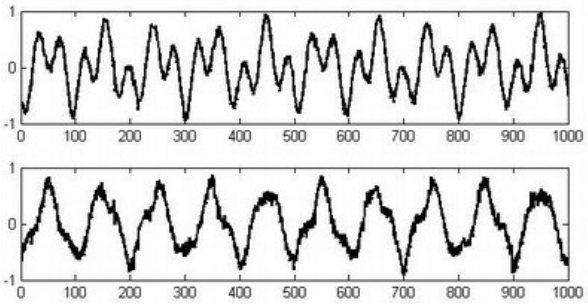

Figura 2.12: *Sinais artificiais misturados.*

A mistura foi efetuada com a matriz disposta na equação (2.2). Esta matriz foi gerada aleatoriamente, apenas para mostrar o processo de mistura com os sinais originais. Após a mistura, os sinais misturados irão compor as linhas da matriz X.

$$A = \begin{bmatrix} -0,6 & 0,5 \\ 0,2 & 0,7 \end{bmatrix}. \tag{2.2}$$

Através da ICA, as fontes puderam ser estimadas, produzindo os sinais apresentados na Figura 2.13. Convém notar que o sinal e a ordem dos sinais originais estimados foram diferentes dos sinais originais, fato este que será explicado nas seções seguintes.

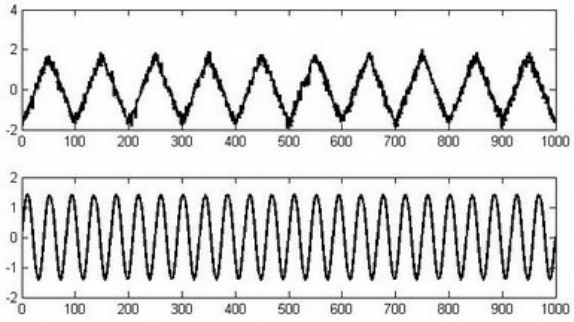

Figura 2.13: *Sinais artificiais desmisturados usando ICA.*

2.5.1 Extração de características usando ICA

Segundo Theodoridis e Koutroumbas (2009), reconhecimento de padrões corresponde à classificação de objetos em categorias ou classes baseada em características. Estas características são extraídas a partir do conjunto de dados usando métodos tais como PCA, ICA, Transformada Discreta de Fourier (DFT), Transformada Discreta do Cosseno (DCT), dentre outras.

Algumas pesquisas tem utilizado ICA para extração de características e reconhecimento de padrões com bons resultados. A razão para isso pode estar nos resultados obtidos ao aplicar ICA para aprender características de imagens naturais a partir de um conjunto de funções base (OLSHAUSEN; FIELD, 1996), e mais tarde aprender características de imagens diversas tais como faces, objetos desenvolvidos pelo homem e até mesmo padrões de textura (BARTLETT; MOVELLAN; SEJNOWSKI, 2002) (CAVALCANTE et al., 2009) (JENSSEN; ELTOFT, 2003).

No contexto da extração de características usando ICA, consideramos X como um conjunto de imagens mamográficas formadas por um conjunto de imagens base A, ou funções base, ponderadas por um conjunto de coeficientes S. Na Figura 2.14, pode-se observar a representação de uma região de uma imagem mamográfica contendo massas representadas como uma combinação linear de suas funções base, que são as colunas da matriz A.

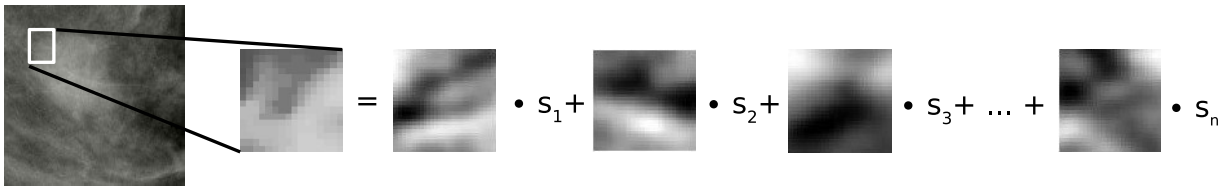

Figura 2.14: *Região de mamografia contendo uma massa representada como uma combinação linear de um conjunto de funções base.*

Para determinar A, deve-se encontrar uma transformação linear W a partir dos dados contidos em X, de modo que o vetor $Y = WX$ seja a estimativa para o vetor de coeficientes S. Quando $A = W^{-1}$ é determinado, as funções base podem ser usadas como filtros para detectar características de massas mamárias em mamografias. Considera-se as funções base como filtros porque têm uma relação convolutiva com os coeficientes correspondentes para formar a imagem (ou parte disso).

2.5.2 Definições

Sejam dadas observações de n sinais misturados, modelados como combinações lineares de n funções base

$$x_i = a_{i1}s_1 + a_{i2}s_2 + a_{ai3}s_3 + \cdots + a_{in}s_n, \tag{2.3}$$

onde cada sinal x_i, bem como cada componente independente s_i, é uma variável aleatória. Em notação matricial, tem-se $X = AS$. O objetivo deste modelo é permitir que se possa estimar o conjunto de funções base em A, bem como a matriz de coeficientes S, somente observando X.

A estimação das componentes é baseada em algumas pressuposições, a saber:

- as componentes independentes s_i são estatisticamente independentes;
- as componentes possuem distribuição não-gaussianas.

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

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

Tais ambiguidades se devem ao fato de A e S serem desconhecidas. Como consequências destas ambiguidades, não é possível determinar as energias ou as amplitudes dos sinais, nem tão pouco os sinais ou a ordem de s_i.

2.5.3 Independência estatística e não-correlação de variáveis

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

$$P(x_1, x_2) = P(x_1) \cdot P(x_2), \tag{2.4}$$

ou em outras palavras, a probabilidade conjunta de x_1 e x_2 é igual ao produto das densidades marginais $P(x_1)$ e $P(x_2)$.

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

$$cov_{x_1,x_2} = E[(x_1 - \mu_1) \cdot (x_2 - \mu_2)] = 0, \tag{2.5}$$

sendo μ_1 e μ_2 as médias das variáveis x_1 e x_2, respectivamente, e $E[\cdot]$ o operador esperança matemática.

2.5.4 Estimação das componentes independentes

A estimação das componentes independentes pode ser obtida através do vetor de funções base, ou matriz de mistura, A, da seguinte forma:

$$\hat{S} = WX, \tag{2.6}$$

de forma que, ao fazer $W = A^{-1}$, teremos $\hat{S} = S$ e portanto, obtemos a estimativa das componentes independentes a partir apenas de X.

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

$$y = b^T X, \tag{2.7}$$

sendo $X = AS$, pode-se escrever

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

onde b deve ser determinado. A partir da equação 2.8, pode-se observar que y é uma combinação linear se s_i, com coeficientes dados por $q = b^T A$. Desta forma, obtemos:

$$y = q^T S. \tag{2.9}$$

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

Uma forma de determinar b é variar os coeficientes em q e verificar como a distribuição de $y = q^T S$ muda. Como consequência do Teorema do Limite Central (PAPOULIS A.; PILLAI, 2002), a soma de variáveis aleatórias aproxima-se cada vez mais de uma distribuição normal do que as variáveis originais, y se aproxima mais de uma variável aleatória gaussiana que qualquer uma das s_i e menos quando comparada a uma das s_i. Assim, apenas um elemento q_i de q é diferente de zero. Como, na prática, os valores de q são desconhecidos, e através das equações 2.7 e 2.9 tem-se que

$$b^T X = q^T S, \tag{2.10}$$

podendo-se variar b e observar a distribuição de $b^T X$.

Dessa forma, pode-se tomar como b um vetor que maximiza a não-gaussianidade de $b^T X$, sendo $q = A^T S$, contendo apenas uma de suas componentes diferente de zero. Isso significa que y na equação 2.7 é igual a uma das componentes independentes, e a maximização da não-gaussianidade de $b^T X$, permite encontrar uma das componentes.

2.5.5 Negentropia como medida de não-gaussianidade

A entropia de uma variável aleatória está relacionada com a quantidade de informação que essa variável possui, sendo maior quanto mais for imprevisível a variável. Em se tratando de variáveis aleatórias, é denominada entropia diferencial. Se y é um vetor

aleatório com função densidade de probabilidade $f(y)$, a sua entropia diferencial é dada por:

$$H(y) = -\int f(y) \log f(y) dy. \tag{2.11}$$

Sabendo-se que uma variável gaussiana tem a maior entropia dentre todas as variáveis aleatórias de igual variância, tem-se que uma versão modificada da entropia diferencial pode ser usada como medida de não-gaussianidade. Tal medida é denominada negentropia, definida por

$$J(y) = H(y_{gauss}) - H(y), \tag{2.12}$$

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

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

$$J(y) \approx \frac{1}{12} E[y^3]^2 + \frac{1}{48} kurt(y)^2, \tag{2.13}$$

sendo $kurt(y)$ a curtose de y, definida como o momento de quarta ordem da variável aleatória y, expressa por

$$kurt(y) = E[y^4] - 3(E[y^2])^2. \tag{2.14}$$

2.6 Análise de Componentes Principais

A análise de componentes principais (PCA) ou transformada de Karhunen-Loève, como também é conhecida, é um dos métodos mais populares para a geração de características em reconhecimento de padrões. Também é largamente usada em aplicações tais como redução de dimensionalidade, compressão de dados com perda e visualização de dados. Pode-se definir PCA como sendo uma projeção ortogonal dos dados em um espaço linear de menor dimensão (BISHOP, 2006).

As componentes principais correspondem a vetores ortogonais sobre os quais os

dados são projetados, seguindo um critério de maior variabilidade. O primeiro vetor representa a maior variância, o segundo vetor representa a segunda maior variância, e assim sucessivamente (COSTA, 2012). Pode-se observar na Figura 2.15(a) um exemplo de distribuição de dados em relação às variáveis x_1 e x_2. Na Figura 2.15(b) o resultado da projeção dos dados nas componentes principais y_1 e y_2. O vetor y_1 representa a direção de maior variância, o vetor y_2, a segunda maior variância e ao mesmo tempo ortogonal a y_1. Esse processo continua, caso hajam mais variáveis.

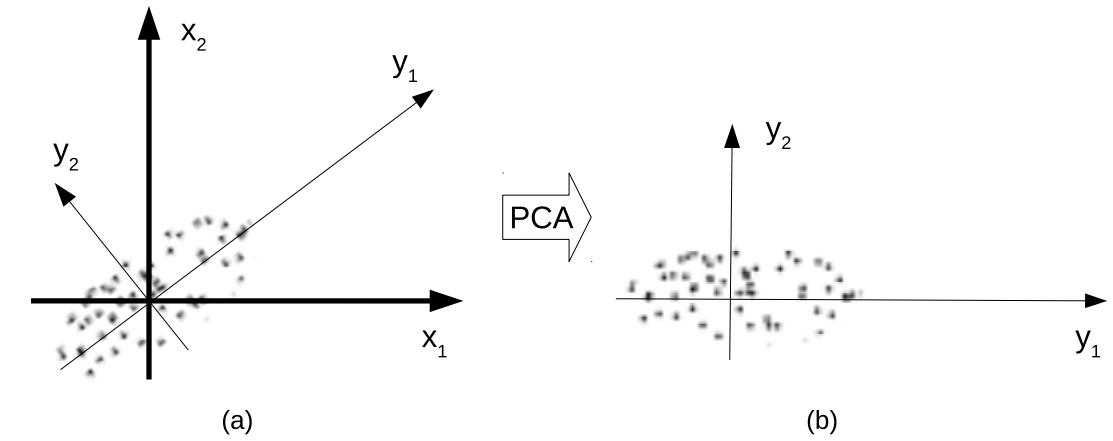

Figura 2.15: *Exemplo de projeção ortogonal com PCA. (a) Distribuição de dados em relação às variáveis x_1 e x_2. (b) Resultado da projeção dos dados nas componentes principais y_1 e y_2.*

Em PCA, os dados $X = [x_1, x_2, ..., x_M]^T$ são a combinação de características mutuamente não correlacionadas. Neste contexto, consideramos X como um conjunto de amostras extraídas da imagem da mama e nós definimos o vetor transformado Y como

$$Y = A^T X. \tag{2.15}$$

As colunas de $A_{M\times M}$ são chamadas vetores base da transformação. Os elementos de Y são as projeções dos elementos de X nestes vetores base, de modo que os componentes de Y não estejam correlacionados. Para obter esses componentes, estima-se a matriz de covariância Σ, onde o valor médio é assumido como zero, $E[X] = 0$. Nesse caso, as matrizes de covariância e autocorrelação coincidem, $R = E[XX^T] = \Sigma$. Dadas M características, x_i, $i = 1, \ldots, M$, a estimativa da matriz de autocorrelação é

$$R = \frac{1}{M}\sum_{i=1}^{M} x_i x_i^T \tag{2.16}$$

Assim, calcula-se o autovalor λ_i e o autovetor a_i e organiza-se os autovalores na ordem decrescente $\lambda_0 \geq \lambda_1 \geq \cdots \lambda_M$. São escolhidos os m maiores autovalores $\lambda_0, \lambda_1, \cdots, \lambda_{m-1}$ e seus respectivos autovetores a_i, $i = 0, 1, 2, \cdots, m-1$, tais que $A = [a_0 \quad a_1 \quad a_2 \quad \cdots \quad a_{m-1}]$.

A matriz A transforma cada vetor x_i, $i = 1, \ldots, M$ no espaço original para um vetor m-dimensional y via transformação $y = A^T x$. As colunas de A são filtros para detectar características de massas estimadas usando PCA.

2.7 Análise de Agrupamentos

Uma técnica de estatística multivariada muito aplicada em análise e reconhecimento de padrões denomina-se análise de agrupamentos. Seu objetivo é agregar objetos com base nas características que eles possuem (HAIR JR. et al., 2009). Existem vários algoritmos de agrupamento, classificados como hierárquicos ou não hierárquicos. Os algoritmos hierárquicos criam subgrupos a partir de grupos e os não hierárquicos apenas classificam os dados usando uma única partição dos dados.

Neste trabalho, aplicou-se um método hierárquico denominado dendrograma. Um dendrograma é um gráfico que retrata o processo de agrupamento (HAIR JR. et al., 2009), representado por uma estrutura em árvore. A idéia principal é criar, primeiramente, agrupamentos entre amostras mais próximas, de acordo com uma medida de distância. A seguir, novos grupos são definidos a partir dos agrupamentos mais próximos criados na etapa anterior. Esse processo continua até que todas as amostras pertençam a um grupo.

Sejam as observações $x = [x_1, x_2, \cdots, x_p]$ e $y = [y_1, y_2, \cdots, y_p]$. Algumas das medidas de distância mais usuais envolvendo os elementos de x e de y são apresentadas na Tabela 2.1.

Tabela 2.1: *Medidas de distância mais usuais em análise de agrupamentos*

Distância	Medida
Manhattan	$\lvert x_1 - y_1\rvert + \lvert x_2 - y_2\rvert + \cdots + \lvert x_p - y_p\rvert$
Euclideana	$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$
Minkowski	$\sqrt[s]{(x_1 - y_1)^s + (x_2 - y_2)^s + \cdots + (x_p - y_p)^s}, s \geq 1$
Chebyshev	$max(\lvert x_1 - y_1\rvert, \lvert x_2 - y_2\rvert, \cdots, \lvert x_p - y_p\rvert)$

O procedimento de agrupamento hierárquico ocorre da seguinte forma:

1. Após a determinação da medida de distância a ser empregada, obtem-se então

a matriz simétrica de distâncias D, definida por:

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix},$$

onde d_{ij} é a distância entre o objeto i e o objeto j, e $d_{11} = d_{22} = \cdots = d_{nn} = 0$.

2. Faz-se um grupo para cada objeto.

3. A partir da matriz D, encontra-se o pares de grupos com menor distância e junta-se estes grupos.

4. Agrupa-se os grupos formados na etapa anterior e recalcula-se as distâncias desse grupo para todos os objetos.

5. Repete-se os 3 e 4 até sobrar um único grupo.

Ao fim deste processo, é possível representar graficamente a formação de cada um dos grupos, e o gráfico gerado corresponde ao dendrograma.

A título de exemplo, seja o conjunto de observações $x = \{2; 4; 7; 10; 11\}$, adotando-se a medida de distância euclideana, tem-se a seguinte matriz de distâncias:

$$D = \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 5 & 3 & 0 & & \\ 8 & 6 & 3 & 0 & \\ 9 & 7 & 4 & 1 & 0 \end{bmatrix}.$$

Percebe-se que os objetos mais próximos são o quarto e o quinto, pois $d_{54} = 1$ é a menor distância. Agrupando-se estes objetos, e calculando-se as novas distâncias substituindo-se os elementos agrupados por suas médias, tem-se a nova matriz de distâncias:

$$D = \begin{bmatrix} 0 & & & \\ 2 & 0 & & \\ 5 & 3 & 0 & \\ 8.5 & 6.5 & 3.5 & 0 \end{bmatrix}.$$

Percebe-se que os objetos mais próximos são o primeiro e o segundo, pois $d_{21} = 2$ é a menor distância. Agrupando-se estes objetos, e calculando-se as novas distâncias substituindo-se os elementos agrupados por suas médias, tem-se a nova matriz de distâncias:

$$D = \begin{bmatrix} 0 & & \\ 4 & 0 & \\ 7.5 & 3.5 & 0 \end{bmatrix}.$$

Percebe-se que os objetos mais próximos são o terceiro e o segundo, pois $d_{32} = 3.5$ é a menor distância. Agrupando-se estes objetos, resta agrupar os dois ultimos objetos restantes, substituindo-se estes objetos por suas médias.

Como resultado, faz-se o esboço gráfico dos agrupamentos efetuados, obtendo-se a representação apresentada na Figura 2.16, denominado dendrograma. O eixo horizontal deste gráfico representam os valores agrupados e o eixo vertical as distâncias calculadas.

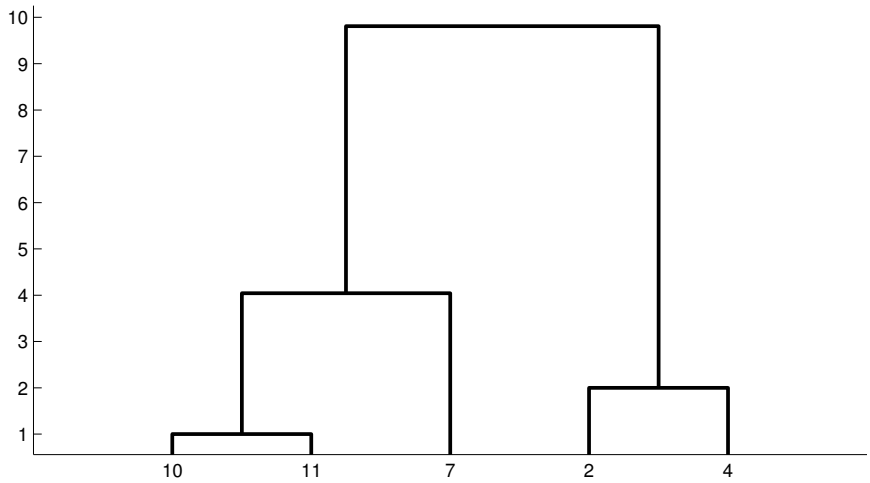

Figura 2.16: *Dendrograma: esboço do agrupamento hierárquico efetuado sobre um conjunto de dados.*

A decisão sobre o agrupamento final é chamada corte do dendrograma. O corte do dendrograma é uma linha traçada na horizontal por todo o dendrograma para especificar os agrupamentos finais. Na figura 2.17 pode-se ver o dendrograma obtido anteriormente, com uma linha de corte definindo dois grupos finais.

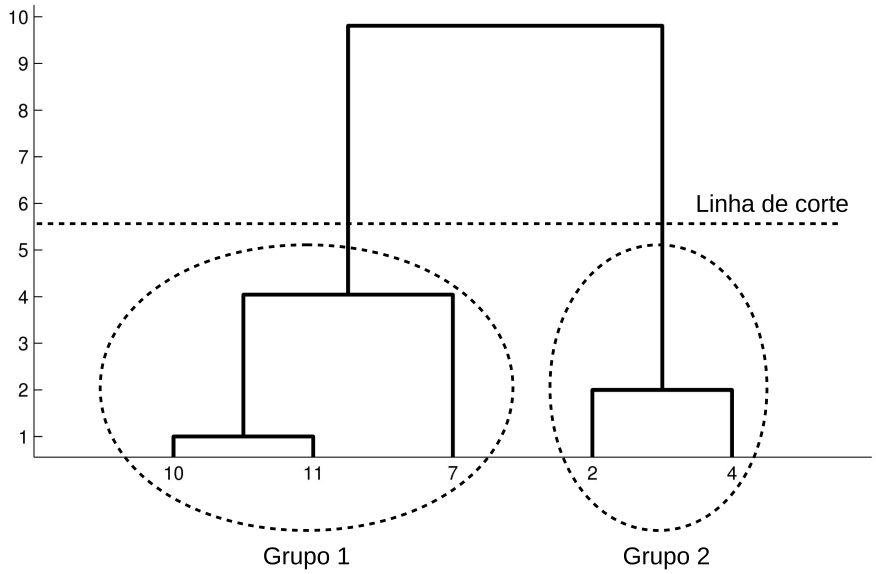

Figura 2.17: *Dendrograma: Linha de corte do dendrograma para determinação de dois agrupamentos finais.*

2.8 K-médias

Trata-se de uma técnica de aprendizagem de máquina não supervisionada, aplicada em problemas de agrupamento de dados. O algoritmo segue um simples e fácil caminho para classificar um determinado conjunto de dados, dado o número k de grupos, sendo k previamente conhecido (THEODORIDIS; KOUTROUMBAS, 2009).

A idéia principal é definir k centróides, um para cada grupo. Cada um dos pontos do conjunto de dados é associado ao k-ésimo centróide mais próximo, usando uma medida de distância pré-estabelecida. Após todos os pontos serem associados, são estabelecidos k novos centróides como sendo centros dos agrupamentos estabelecidos na etapa anterior. Nesse momento, cada ponto é novamente associado ao k-ésimo centróide mais próximo e assim por diante, até que não haja mais mudança na localização dos k centróides. Na Figura 2.18, é possível ver um exemplo do funcionamento do k-médias. Na Figura 2.18(a), tem-se as amostras aleatórias e os centroides em suas localizações iniciais. Na Figura 2.18(b), os centroides em suas posições definitivas e os grupos a que as amostras pertencem.

Na formulação clássica, o objetivo do algoritmo k-médias é minimizar a função objetivo J que representa o erro quadrático médio entre um ponto de dados $r(z, y)$ e o i-ésimo centróide c_k, de acordo com a equação

$$J = |r(z, y) - c_k|^2, \tag{2.17}$$

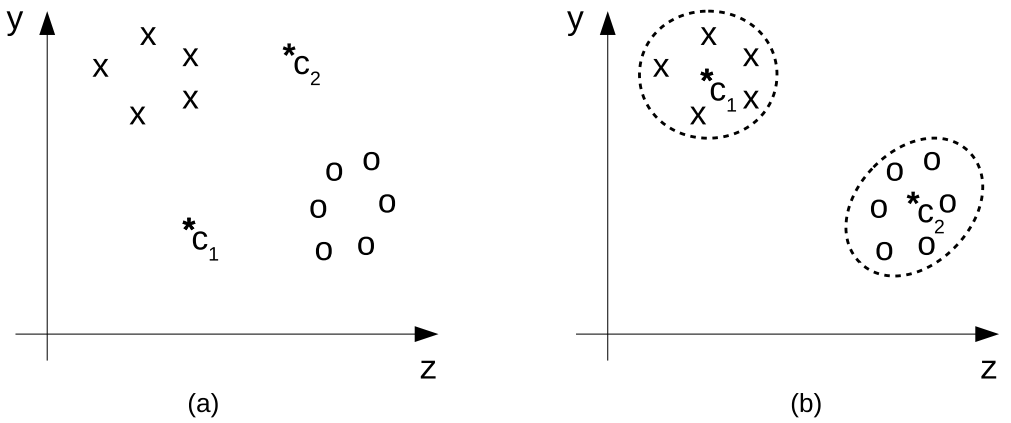

Figura 2.18: *Exemplo do funcionamento do k-médias. (a) Amostras aleatórias e os centroides em suas localizações iniciais. (b) Centroides em suas posições definitivas e os grupos a que as amostras pertencem.*

onde $|r(z,y) - c_k|^2$ é a distância entre $r(z,y)$ e c_k, e indica a similaridade entre cada ponto de dado e seu respectivo grupo. Assim, todos os pixels são atribuídos ao centro mais próximo com base nessa distância. Depois de todos os pixels terem sido atribuídos, os centros são recalculados usando,

$$c_k = \frac{1}{k} \sum_{r(z,y) \in S_k} r(z,y), \tag{2.18}$$

onde S_k é um conjunto de pixels associados com c_k. Esse processo é repetido até os centroides não se moverem, produzindo um grupo de pixels com base na similaridade.

Capítulo 3

Materiais e Métodos

Primeiramente foram obtidas as imagens usadas no processo. Para os métodos, dois grupos de regiões de interesse (ROIs) foram obtidos a partir de duas bases de dados públicas: a base de dados de mamografia digital da sociedade de análise de imagem mamográfica Mini-MIAS (SUCKLING et al., 1994) e o banco de dados digital para rastreio de mamografia DDSM (HEATH et al., 2001). Para cada ROI obtida de mama não densa que contém massa, aproximadamente 100 amostras aleatórias foram extraídas das regiões de *ground truth* fornecidas pelos bancos de dados. Essas amostras têm dimensão de 16x16 pixels. Cada amostra foi reorganizada como um vetor coluna e organizada na matriz de entrada dos algoritmos de treinamento usados para obter as funções base, que serão usadas como filtros para detectar as massas nas imagens de teste. Os filtros foram aplicados nas imagens de teste e as respostas dos filtros foram adicionadas e agrupadas usando o algoritmo k-médias. Após o agrupamento, foi realizada uma busca pelo grupo com maior energia, que corresponde à região de massa detectada pelos filtros. O diagrama do método proposto é mostrado na Figura 3.1.

Também foi feita a seleção dos melhores filtros para caracterizar regiões de massa nas imagens. Os filtros selecionados foram usados na etapa de teste, para indicar a presença ou ausência de massas.

3.1 Aquisição das imagens

O banco de dados Mini-MIAS possui imagens de mamografia de 322 pacientes, digitalizadas em uma resolução de 200 mícrons e foram redimensionadas para 1024 x 1024 pixels, 8 bits. O banco de dados Mini-MIAS fornece um registro contendo informações dos casos, tais como tipo de tecido, classe da anomalia, tipo de anomalia (benigna ou maligna)

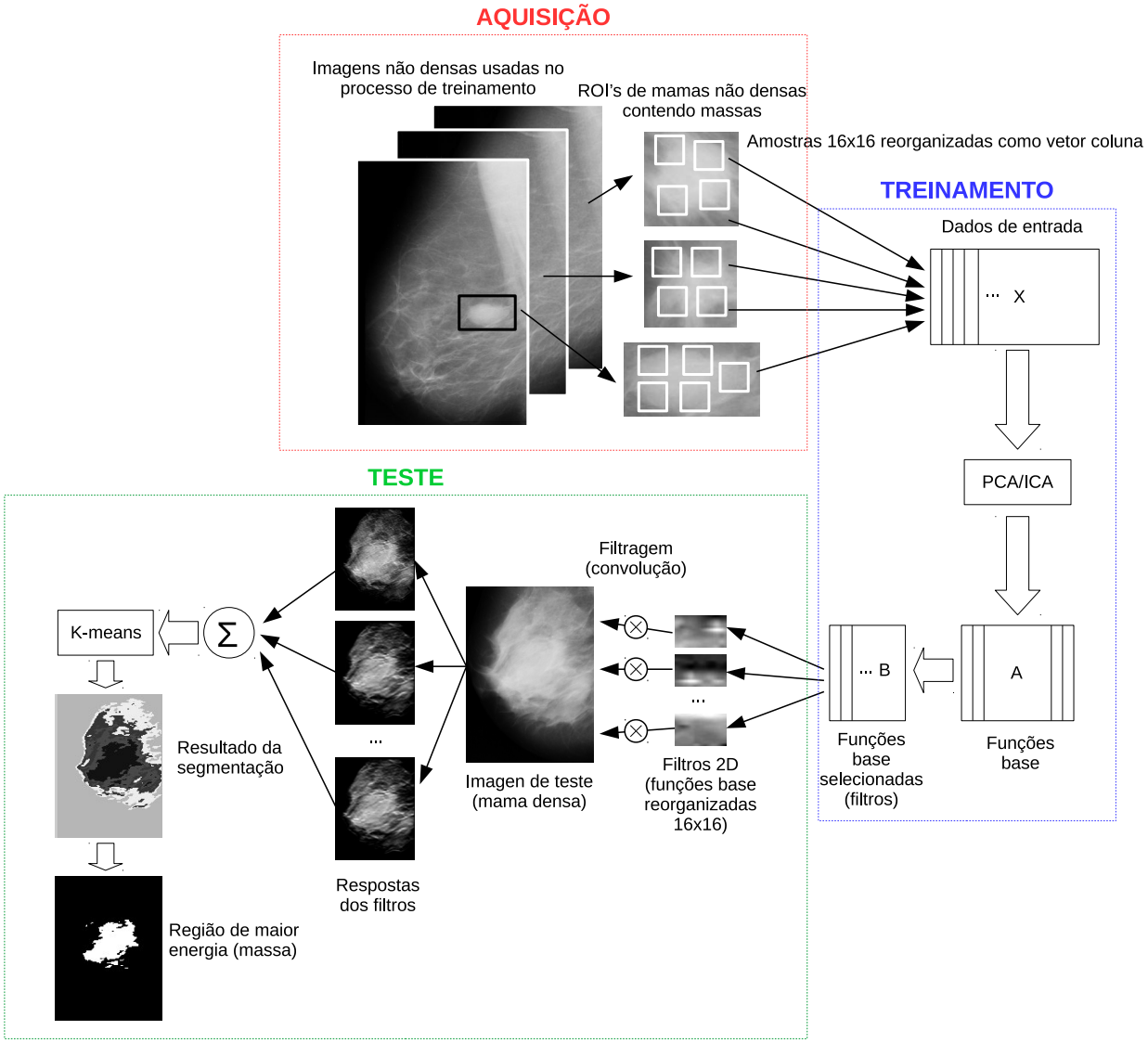

Figura 3.1: *Diagrama do método proposto. Um conjunto de imagens não densas que contém lesão é usado no processo de treinamento. A área da mama que contém a lesão (ROI's) é selecionada e as amostras de 16x16 pixels, sem sobreposição de amostras, são extraídas, reorganizadas como um vetor coluna e concatenadas. Este vetor é usado nos algoritmos de treinamento para extrair funções base, a partir dos quais os melhores filtros são selecionados. As respostas dos filtros são somadas e agrupadas usando k-médias. A partir do resultado da segmentação, selecionou-se o grupo com a maior energia, que corresponde à região de massa.*

e as coordenadas do centro da anomalia, bem como o raio aproximado (em pixels) de um círculo que circunda a anomalia. Neste trabalho, consideramos esse círculo como *ground truth* para a avaliação da região detectada pelo método em relação ao especificado nos registros do banco de dados Mini-MIAS.

O banco de dados DDSM tem aproximadamente 2400 casos, digitalizados em uma resolução que varia entre 42 e 50 mícrons, salvos em compressão LOSSLESS JPEG, ou seja, compressão com perda, 16 bits. O banco de dados DDSM fornece um registro contendo a densidade, tipo e localização da lesão, sob a forma de um código de cadeia (GON-

ZALEZ; WOODS; EDDINS, 2003). Para manter a homogeneidade do conjunto de dados nos experimentos, as imagens do banco de dados DDSM foram convertidas para 8 bits e redimensionadas para 1024 x 1024. Assim, para treinamento e testes, as imagens usadas na base Mini-MIAS e DDSM têm a mesma resolução e tamanho.

Nos experimentos realizados, foram usadas imagens de mamas não densas e mamas densas. A densidade da mama é estimada de acordo com a escala BI-RADS. Na Figura 3.2, é possível observar dois exemplos de imagens usadas no experimento, na Figura 3.2(a), mama densa sem massa e na Figura 3.2(b), mama densa com massa e sua localização, mostrados de acordo com o registro do respectivo banco de dados.

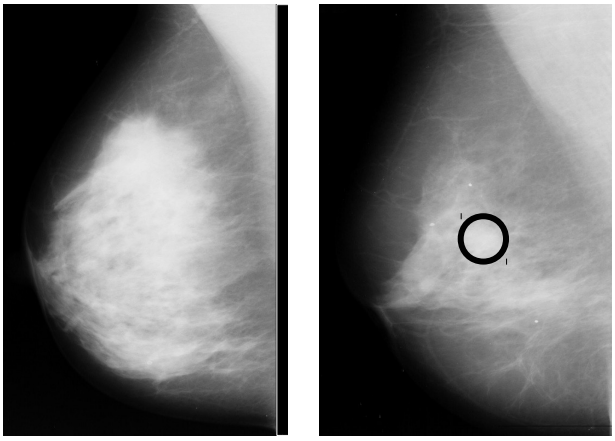

Figura 3.2: *Exemplos de imagens de mamografia usadas nos experimentos. (a) mama densa sem massa. (b) mama densa com massa e sua localização correspondente são mostrados de acordo com o registro do banco de dados. Fonte: Suckling et al. (1994).*

Para o treinamento, 36 ROIs contendo lesão foram usadas do banco de dados Mini-MIAS e 83 ROIs contendo lesão foram usadas do banco de dados DDSM, todas obtidas a partir de mamas não densas. Não foram utilizados ROIs contendo microcalcificações nos experimentos. Na Figura 3.3, é possível observar algumas ROIs utilizadas para no treinamento do algoritmo.

Para os testes, foram utilizadas 56 ROIs contendo lesão do banco de dados Mini-MIAS e 83 ROIs contendo lesão do banco de dados DDSM, todas obtidas a partir de mamas densas. Para cada ROI, uma imagem binária de *ground truth* foi gerada, de acordo com os registros do banco de dados correspondente. Essas imagens de *ground truth* foram utilizadas no processo de avaliação da segmentação e de detecção de lesões. Na Figura 3.4, observa-se algumas ROIs usadas nos testes e suas respectivas imagens de *ground truth*.

As imagens usadas no treinamento, de acordo com os registros do banco de dados Mini-MIAS contém o atributo tecido de fundo igual a *Fatty*, o que caracteriza imagens de mamas não densas. Além disso, as imagens que contém os seguintes valores para o atributo

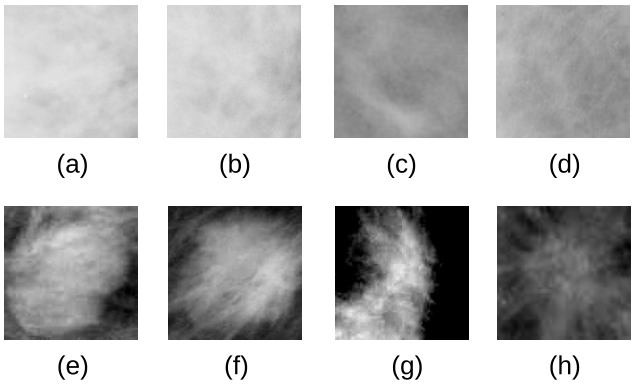

Figura 3.3: *Exemplos de ROIs utilizados no treinamento do algoritmo. (a)-(d) regiões de mamas densas sem massa. (e)-(h) regiões de mamas não densas contendo massas.*

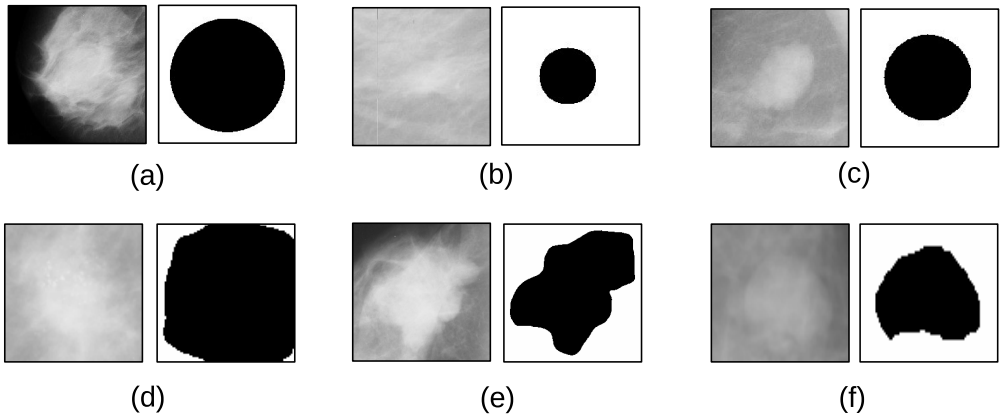

Figura 3.4: *Exemplos de ROIs utilizados nos testes e suas respectivas imagens de ground truth: (a) caso mdb001 (MIAS); (b) caso mdb013 (MIAS); (c) caso mdb015 (MIAS); (d) caso A_1025_1-2.LEFT_CC (DDSM); (e) caso A_1027_1-1.LEFT_CC (DDSM); (f) caso B_3133_1-1.RIGHT_MLO (DDSM).*

classe de anormalidade foram selecionadas: assimetria, massas circunscritas, massas mal definidas, distorção arquitetural e massas espiculadas. Esses valores caracterizam imagens com lesões de diferentes tipos, exceto microcalcificações. Para os testes, os valores para o atributo classe de anormalidade foram os mesmos, mas para o atributo tecido de fundo, foram selecionadas imagens que tiveram os valores *Fatty-glandular* e *Dense-glandular*, que compõem as imagens de mamas densas do banco de dados.

As imagens do banco de dados DDSM estão disponíveis em volumes compactados, distribuídos da seguinte forma: 12 volumes com 695 imagens sem lesões, 14 volumes com 870 imagens contendo lesões benignas e 15 volumes com 915 imagens contendo lesões malignas. As imagens usadas para treinamento foram obtidas a partir dos volumes benign_01, benign_02 e cancer_01, cancer_02. A partir das imagens disponíveis nesses volumes, foram utilizados apenas casos com densidade 1 e 2, que correspondem às imagens da densidade I e II, na escala BI-RADS. Para os testes, utilizaram-se os volumes benign_03, benign_04, cancer_03, cancer_04 e normal_01. Apenas as imagens com

densidade 3 e 4 foram utilizadas, que correspondem às imagens da densidade III e IV, na escala BI-RADS.

3.2 Extração de características

Reconhecimento de padrões corresponde à classificação de objetos em categorias ou classes com base em características (THEODORIDIS; KOUTROUMBAS, 2009). Essas características são extraídas do conjunto de dados usando métodos tais como análise de componentes principais (PCA), ICA, transformada discreta de Fourier (DFT), transformar discreta coseno (DCT), entre outros.

ICA vêm sendo amplamente utilizada em processamento de sinal e imagem (HYVARINEN, 1999) (HYVARINEN; OJA, 2000) (HYVARINEN; KARHUNEN; OJA, 2004). Algumas dessas aplicações envolvem processamento de informações biomédicas, tais como sinais de fMRI, EEG e ECG (JAMES; HESSE, 2005). Além disso, ele tem sido aplicado em outras áreas para tarefas como a remoção de ruído (VOROBYOV; CICHOCKI, 2002), previsão de séries temporais para a bolsa de valores (MOK; LAM; NG, 2004), reconhecimento de voz (CHIEN; CHEN, 2006) e extração de características para o reconhecimento de padrões (HOYER; HYVARINEN, 2000; OLSHAUSEN; FIELD, 1996; BARTLETT; MOVELLAN; SEJNOWSKI, 2002; CAVALCANTE et al., 2009; JENSSEN; ELTOFT, 2003).

3.2.1 Treinamento usando ICA

Uma vez que as amostras foram obtidas e dispostas na matriz de entrada X, os filtros para detectar características das massas foram estimados usando ICA. Como apresentado na equação 2.2, a matriz A é obtida e encontramos nas colunas da matriz A as funções base que formaram as imagens de entrada na matriz X. Neste trabalho, utilizamos o algoritmo FastICA (HYVARINEN, 1999) devido à sua simplicidade e convergência rápida (HYVARINEN, 2017).

Foram extraídas as funções base de ROIs normais densas e de ROIs não densas contendo massas. Na Figura 3.5(a), pode-se observar exemplos de funções base do tecido normal não denso, e na Figura 3.5(b), as funções base para o tecido lesionado não denso. Pode-se ver na Figura 3.6(a), e na Figura 3.6(d), um conjunto de funções base derivadas de ROIs densas sem massa e ROIs não densas com massa, respectivamente. Na Figura 3.6(b) e na Figura 3.6(e), pode-se ver as médias e variâncias de cada um dos conjuntos de funções base e na Figura 3.6(c), e na Figura 3.6(f), seus respectivos espectros de Fourier.

É possível observar a diferença entre os dois espectros.

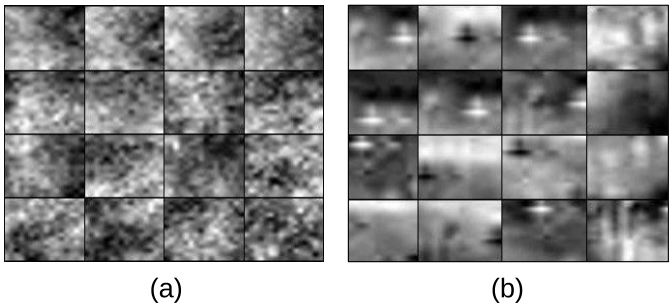

Figura 3.5: *Exemplo de funções base, obtidas com ICA: (a) funções base de tecido normal não denso; (b) funções base de tecido lesionado denso.*

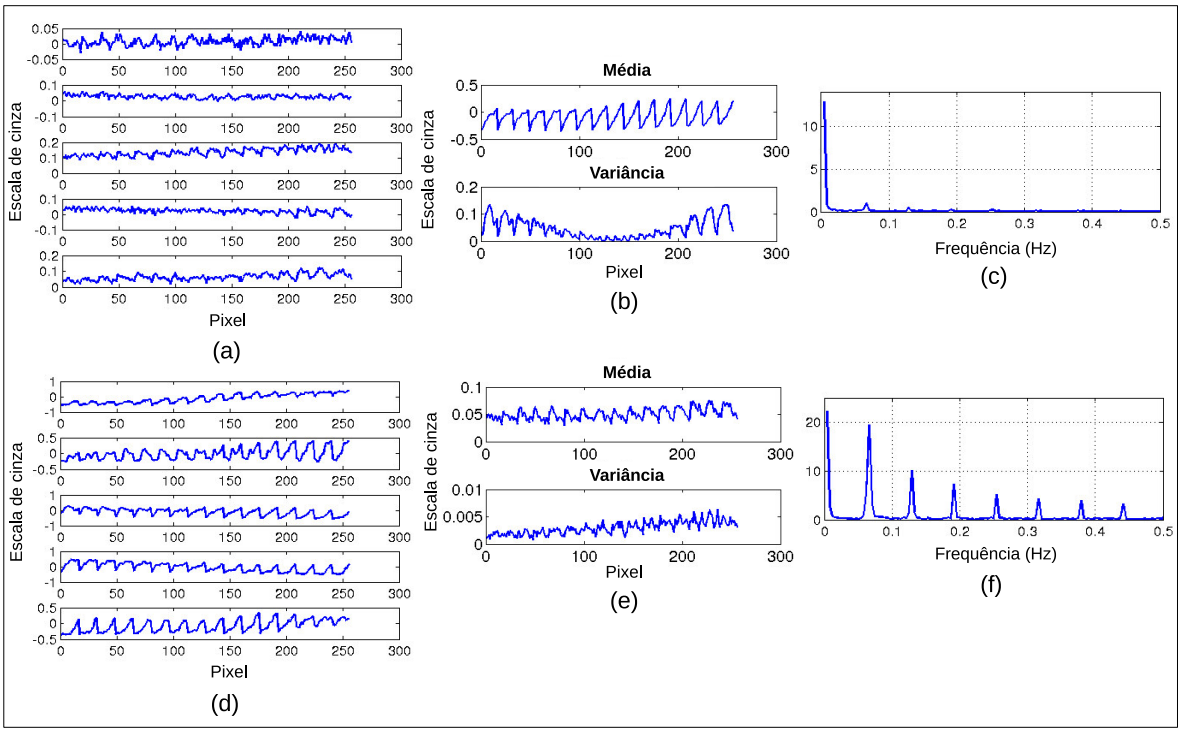

Figura 3.6: *Exemplo de um conjunto de funções base representadas como vetores coluna e seu correspondente espectro de Fourier. (a) Algumas funções base de regiões densas sem massa. (b) Média das funções base e variância das funções base de regiões densas sem massa. (c) Espectro de Fourier da média das funções base de regiões densas sem massa. (d) Algumas funções base de regiões não densas que contêm massa. (e) Média das funções base e variância das funções base de regiões não densas contendo massa. (f) Espectro de Fourier da média das funções base de regiões não densas contendo massa.*

Neste trabalho, foi usando o modelo ICA completo. Desta forma, foi obtida uma matriz quadrada A, com dimensões 256x256, ou seja, com 256 funções base. A seleção das funções base foi realizada por agrupamento do conjuntos de funções base usando k-médias com três grupos.

O número de grupos foi estimado após uma análise de agrupamentos das funções base, utilizando-se o software MATLAB. A determinação dos grupos foi realizada usando-

se como critério as distâncias euclideanas entre as funções base que estivessem dentro do limiar de 70% da maior distância calculada. Pode-se ver na Figura 3.7 o dendrograma feito a partir do conjunto de funções base, onde é possível notar a presença dos três grupos de funções base com características semelhantes. Depois de agrupar as funções base, foram extraídas as funções base médias para cada grupo, a partir do qual calculou-se a curtose e a variância. O grupo de filtros correspondente à maior curtose e maior variância foi utilizado nos experimentos. Nos experimentos, foram selecionadas 53 funções base usando a variância mais alta e 54 funções base usando a maior curtose.

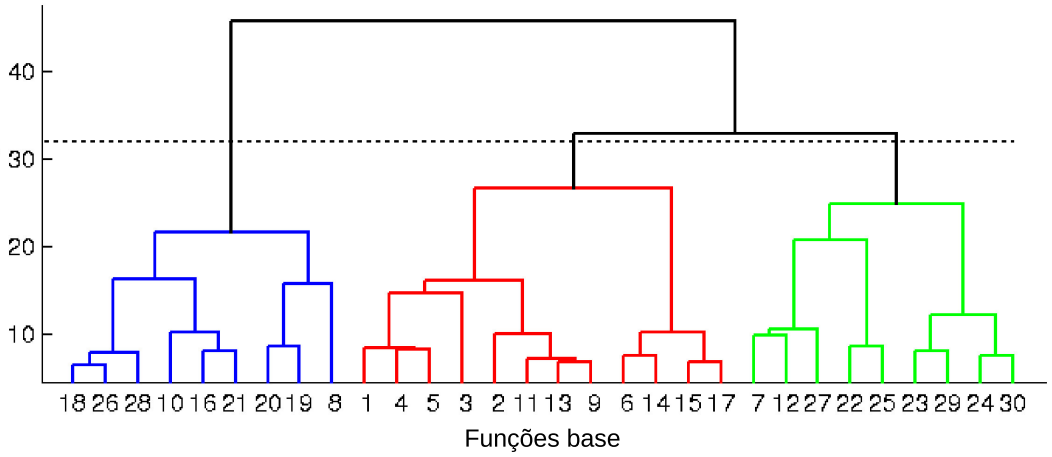

Figura 3.7: *Dendrograma das funções base obtidas com ICA.*

3.2.2 Treinamento usando PCA

Foram extraídas as funções base a partir de ROIs densas normais e de ROIs não densas contendo massas. Pode-se observar na Figura 3.8(a), alguns exemplos de funções base de tecido normal não denso, e na Figura 3.8(b), alguns exemplos de funções base de tecido lesionado não denso.

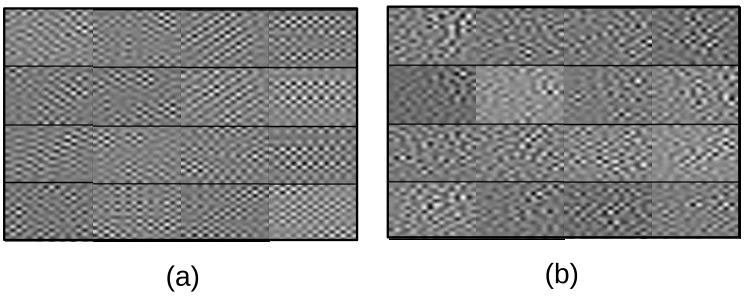

Figura 3.8: *Exemplo de funções base, obtidas de mamas não densas com PCA: (a) funções base de tecido normal. (b) funções base para o tecido lesionado.*

Os filtros foram escolhidos a partir dos autovetores associados aos maiores autovalores e o número de filtros foi definido a partir de um gráfico de Pareto gerado a partir

dos autovalores. Pode-se observar na Figura 3.9 o gráfico de Pareto construído com base nas funções obtidas com PCA. Ao examinar a figura, pode-se perceber que as primeiras oito componentes explicam 90% dos dados. Este foi o número de filtros de PCA usados nos experimentos.

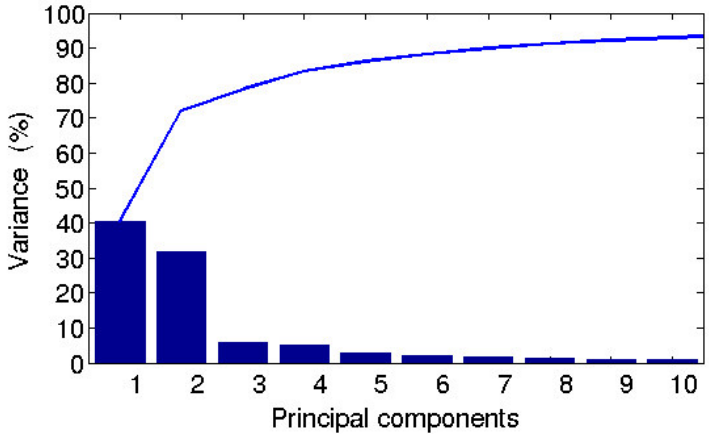

Figura 3.9: *Gráfico de Pareto construído a partir das componentes principais obtidas com PCA. No gráfico, pode-se ver que as oito componentes explicam 90% dos dados.*

3.2.3 Filtragem

Após a obtenção dos filtros utilizando ICA e PCA, os testes foram realizados utilizando ROIs extraídos de mamografias de mamas densas de pacientes com massas benignas ou malignas. De cada uma das imagens de teste, as ROIs foram extraídas, excluindo-se a região do músculo peitoral.

A i-ésima coluna do vetor $\mathbf{a}_i{}^T$, ou i-ésimo filtro selecionado de A, foi usado em uma convolução 2D com imagens de teste $I(z, y)$, onde Z e y são o eixo da imagem, obtendo a seguinte resposta,

$$\{I(z,y) \otimes \mathbf{a}_1, I(z,y) \otimes \mathbf{a}_2, ..., I(z,y) \otimes \mathbf{a}_N\}, \tag{3.1}$$

onde o operador $\otimes$ representa a convolução 2D.

Somente as saídas do filtro não são apropriadas para identificar as características de textura. Foram propostos vários métodos de extração de características para extrair informações úteis das saídas dos filtros (JAIN; F., 1991). As não-linearidades mais utilizadas são a função quadrada $F(z, y) = (I(z, y) \otimes \mathbf{a}_1)^2$ e a função magnitude $F(z, y) = |I(z, y) \otimes \mathbf{a}_1|$. Combinamos essas duas funções para a extração das características $F(z, y) = |I(z, y) \otimes \mathbf{a}_1|^2$.

As respostas aos filtros foram somadas para formar o resultado final,

$$R(z,y) = |I(z,y) \otimes \mathbf{a}_1|^2 + |I(z,y) \otimes \mathbf{a}_2|^2 + ... + |I(z,y) \otimes \mathbf{a}_N|^2. \quad (3.2)$$

Esta soma representa a resposta dos filtros utilizados para detectar as características das regiões lesionadas nas imagens de teste e, portanto, resultou em uma imagem filtrada. Após a soma das respostas dos filtros, foi usado o algoritmo de agrupamento k-médias para agrupar os pixels que apresentaram características semelhantes.

3.2.4 Avaliação do diagnóstico

A avaliação do diagnóstico baseou-se no *ground truth* de cada imagem testada, de acordo com os registros do banco de dados Mini-MIAS e DDSM. Consideramos os pixels detectados pelos filtros dentro do *ground truth* como sucessos (verdadeiros positivos) e detectados fora do *ground truth* como erros (falsos negativos). A avaliação do método proposto foi realizada através da análise de acurácia, sensibilidade e especificidade, conforme definido abaixo,

$$\text{Sensibilidade} = \frac{TP}{(TP+FN)}, \quad (3.3)$$

$$\text{especificidade} = \frac{TN}{(TN+FP)}, \quad (3.4)$$

$$\text{Acurácia} = \frac{(TP+TN)}{(TP+FP+TN+FN)}. \quad (3.5)$$

Sensibilidade, obtida da equação 3.3, refere-se à capacidade do método para prever as imagens que realmente possuem alguma massa. Especificidade, calculada a partir da equação 3.4, refere-se à capacidade do método de inferir que as imagens não possuem massas. A acurácia corresponde à proporção da identificação correta do grupo estudado, calculada a partir da equação 3.5.

Para avaliar o desempenho da segmentação obtida em cada imagem, foi usada a Medida de Sobreposição de Área (AOM), também conhecida como medida de similaridade Jaccard, foi usada (POLAK; ZHANG; PI, 2009). Esta métrica considera a região da imagem detectada pelos filtros dentro do *ground truth* definido nos registros dos bancos de dados utilizados. A medida Jaccard é definida por:

$$AOM = \frac{A_{Seg} \cap A_{GT}}{A_{Seg} \cup A_{GT}}, \tag{3.6}$$

onde A_{Seg} são os pixels de segmentação resultantes e A_{GT} são os pixels considerados como *ground truth* em cada caso. O caso ideal ocorre quando o AOM é 1. Esta métrica foi usada para comparar a segmentação dos filtros de ICA com os filtros de PCA. Neste trabalho, um limite definido como 0.1 foi considerado para a tarefa de detecção.

A Figura 3.10 ilustra a relação entre A_{Seg} e A_{GT}, em três casos diferentes. Os retângulos claros ilustram as regiões resultantes da segmentação após o agrupamento, enquanto que os retângulos escuros representam o *ground truth* da imagem de teste. Para cada tamanho da região segmentada, assumindo que a mesma está totalmente inserida no *ground truth*, tem-se um valor resultante para a medida Jaccard.

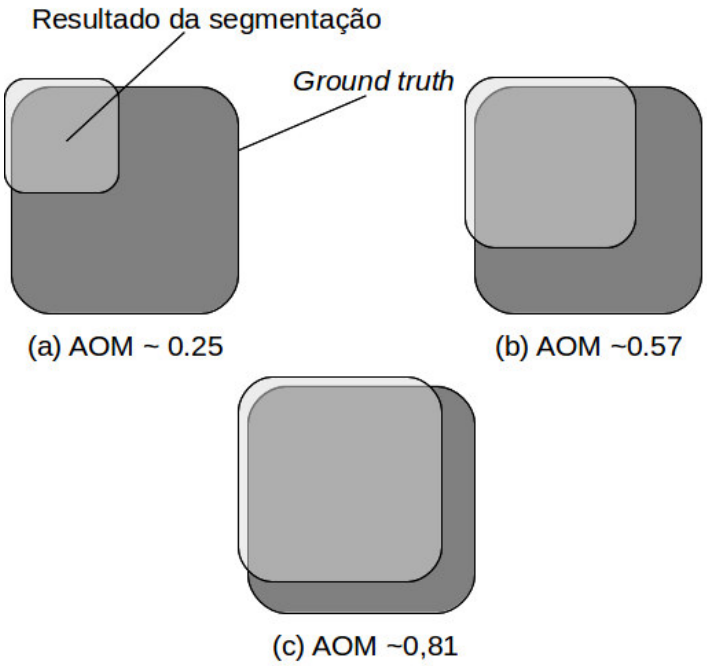

Figura 3.10: *Relação entre A_{Seg} e A_{GT}, em três casos diferentes. Os retângulos claros ilustram o resultado da segmentação, enquanto que os retângulos escuros representam o ground truth da imagem de teste. Abaixo de cada caso, o valor aproximado da métrica de Jaccard.*

3.2.5 Teste Z de hipóteses para duas proporções

Para verificar se existe uma diferença estatistica significante entre os resultados encontrados em mamas densas e mamas não densas, foi efetuado um teste para comparar duas proporções para sensibilidade (ARMITAGE, 2001). Este teste é usado para avaliar uma regra de decisão que permite aceitar ou rejeitar a hipótese de que duas populações tenham a mesma proporção para uma determinada variável, com base em elementos de

amostra. Com este teste, podemos avaliar se há diferença estatística significativa entre os resultados de detecção.

Suponhamos que G e H são variáveis aleatórias independentes que determinam os resultados de experimentos em mamas não densas e em mamas densas, respectivamente. Nesse caso, percebemos que G e H são variáveis Bernoulli com os respectivos parâmetros p_G e p_H. Assim, as hipóteses H_0 e H_1 são estabelecidas como,

$$\begin{cases} H_0 : p_G = p_H \\ H_1 : p_G \neq p_H \end{cases} \tag{3.7}$$

Depois de estabelecido um nível de significância α, definimos a variável padrão Z como sendo

$$Z = \frac{p_G - p_H}{\sqrt{\frac{\rho(1-\rho)}{n_G} + \frac{\rho(1-\rho)}{n_H}}} \sim N(0,1), \tag{3.8}$$

onde Z corresponde à variável padronizada utilizada no teste de significância para duas proporções, n_G e n_H correspondem ao tamanho das amostras de imagens de mamas não densas e mamas densas avaliadas, enquanto p_G e p_H correspondem à proporções diagnósticos corretos, respectivamente. A variável ρ foi calculado por expressão

$$\rho = \frac{n_G p_G + n_H p_H}{n_G + n_H}. \tag{3.9}$$

Capítulo 4

Resultados

Aqui são apresentados os resultados dos dois métodos usados para detectar e segmentar a imagem mamográfica. Nas experiências realizadas, utilizamos o pacote fastICA (HYVARINEN, 2017) escrito em MATLAB e disponível gratuitamente em <https://research.ics.aalto.fi/ica/fastica/>. Os parâmetros utilizados no treinamento FastICA são aqueles definidos como padrão, ou seja, função de não linearidade $g(u) = u^3$, $\mu = 1$, $\epsilon = 0.0001$ e número máximo de iterações para convergência igual Para 1000.

No primeiro experimento, o objetivo é testar quais técnicas poderiam detectar e segmentar melhor a lesão. Os testes foram realizados variando o número de agrupamentos de 4 a 6, com a métrica Jaccard maior ou igual a 0,1. Acima de 6 grupos, os resultados de detecção não foram satisfatórios. Os resultados obtidos na detecção com todos os filtros de ICA e PCA estão na Tabela 4.1. Os resultados obtidos com os filtros de ICA e PCA selecionados usando os critérios de maior curtose são mostrados na Tabela 4.2. Os resultados obtidos com os filtros de ICA e PCA selecionados usando os critérios de maior variância são mostrados na Tabela 4.3.

Tabela 4.1: *Resultados dos testes de detecção em ROIs com todas as funções base de ICA versus PCA (métrica de Jaccard* ≥ 1*).*

	k=4	k=5	k=6
ICA (Mamas densas)	84.71%	79.63%	75.43%
PCA (Mamas densas)	79.94%	72.77%	60.19%
ICA (Mamas não densas)	89.39%	87.58%	83.79%
PCA (Mamas não densas)	85.83%	76.47%	69.90%

No segundo experimento, o objetivo é obter a acurácia dos filtros de ICA na detecção da lesão em mamas densas e não densas. As imagens de mama com e sem lesões

Tabela 4.2: *Resultados dos testes de detecção em ROIs com funções base de ICA de maior curtose versus PCA (métrica de Jaccard* ≥ 1*).*

	k=4	k=5	k=6
ICA (Mamas densas)	84.71%	75.72%	73.02%
PCA (Mamas densas)	82.33%	74.87%	58.40%
ICA (Mamas não densas)	88.19%	86.98%	82.58%
PCA (Mamas não densas)	84.44%	77.07%	70.50%

Tabela 4.3: *Resultados dos testes de detecção em ROIs com funções base de ICA de maior variância versus PCA (métrica de Jaccard* ≥ 1*).*

	k=4	k=5	k=6
ICA (Mamas densas)	85.61%	79.31%	73.91%
PCA (Mamas densas)	81.14%	73.06%	61.39%
ICA (Mamas não densas)	89.39%	86.80%	83.79%
PCA (Mamas não densas)	85.83%	77.07%	77.07%

usadas no experimento foram selecionadas a partir do conjunto de teste obtido dos dois bancos de dados. Para mamas sem lesão, regiões da mama foram tomadas aleatoriamente. Para mamas contendo lesão, foram obtidas ROIs diferentes das utilizadas no primeiro experimento. Desta forma, foram obtidas 60 ROIs contendo lesão e 60 ROIs sem lesão dos dois bancos de dados. Cada uma das 120 ROIs foi filtrada usando as funções base das regiões normais e as funções base das lesões.

Assim, cada imagem testada tinha dois conjuntos de respostas dos filtros. Cada conjunto foi somado separadamente, de acordo com a equação 3.2, e foi produzida uma soma das respostas dos filtros de regiões normais e uma soma das respostas dos filtros de regiões contendo lesão.

Como no primeiro experimento, o algoritmo k-médias foi aplicado à soma das respostas dos filtros de lesão e os pixels pertencentes ao grupo de maior energia foram selecionados e as respectivas posições desses pontos de pixels foram obtidas.

Essas posições marcam a localização dos pixels que correspondem às lesões na imagem original. Os pixels pertencentes a essas posições compõem o vetor P_s. Foi feita uma busca pelos pixels correspondentes a essas posições na soma das respostas dos filtros de região normais. Os pixels pertencentes a essas posições compõem o vetor P_n. A diferença entre os conjuntos P_s e P_n foi calculada e, a partir desta diferença, a variância foi calculada e, para as mamas que contém lesões, o AOM da região segmentada doi calculado.

Na Figura 4.1, é possível observar o resultado do processamento efetuado na imagem A_1388_1_RIGHT_MLO (DDSM). Esta imagem contém uma lesão. Na Figura 4.1(a), pode-se notar a região de interesse, na Figura 4.1(b), o *ground truth* da lesão. Na Figura 4.1(c), o resultado da segmentação, na Figura 4.1(d) a soma das respostas dos filtros de regiões normais, e na Figura 4.1(e), a soma das respostas dos filtros de regiões contendo lesão. Para este caso, a diferença dos somas apresentou variação 0,0105 e a segmentação apresentou AOM 0,1911.

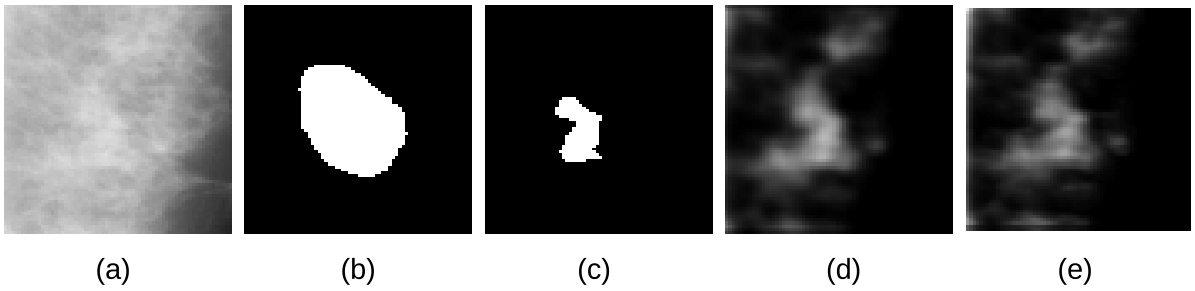

Figura 4.1: *Resultado do processamento para a imagem A_1388_1_RIGHT_MLO (DDSM): (a) região de interesse. (b) ground truth da lesão. (c) Resultado da segmentação. (d) Soma das respostas de filtros de regiões normais. (e) Soma das respostas dos filtros de regiões contendo lesão.*

Na Figura 4.2, pode-se observar o resultado do processamento efetuado na imagem A_0005_1_LEFT_MLO (DDSM). Esta imagem não possui lesão. Na Figura 4.2(a), pode-se notara a região de interesse, na Figura 4.2(b), o resultado da segmentação, na Figura 4.2(c), a soma das respostas dos filtros da regiões normais e na Figura 4.2(d), a soma das respostas dos filtros de regiões contendo lesão. Para este caso, a diferença das somas apresentou variação zero.

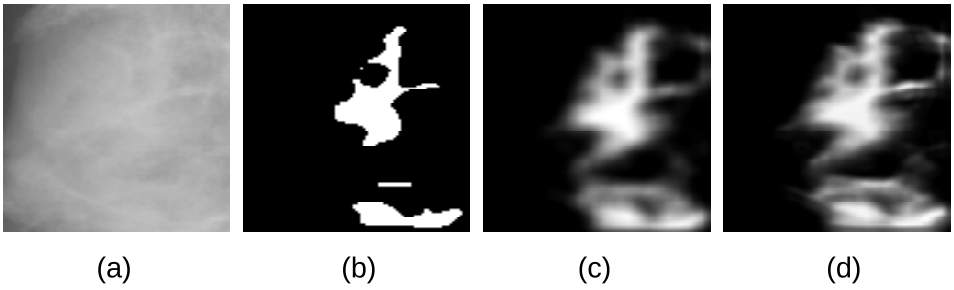

Figura 4.2: *Resultado do processamento para a imagem A_0005_1_LEFT_MLO (DDSM): (a) região de interesse. (b) Resultado da segmentação. (c) Soma das respostas dos filtros de regiões normais. (d) Soma das respostas dos filtros de regiões contendo lesão.*

Após os testes realizados, observou-se que a maioria das imagens de mamas sem lesão utilizadas no experimento apresentaram uma variância para o conjunto P_n igual a zero. Os resultados experimentais em 120 imagens de mamas densas usadas nos testes são mostrados na Tabela 4.4. Nas experiências, a sensibilidade foi de 66,67%, a especificidade foi de 91,67% e a acurácia foi de 79,17%.

Tabela 4.4: *Matriz de confusão para imagens de mamas densas*

	Com massa	Sem massa	Total
Resultados positivos	40 (TP)	5 (FP)	45
Resultados negativos	20 (FN)	55 (TN)	75
Total	60	60	120

Os resultados experimentais obtidos em 96 imagens de teste de mamas não densas são mostrados na Tabela 4.5. Nos experimentos, a sensibilidade foi de 95,65%, a especificidade foi de 90,00% e a acurácia foi de 92,71%.

Tabela 4.5: *Matriz de confusão para imagens de mamas não densas*

	Com massa	Sem massa	Total
Resultados positivos	44 (TP)	5 (FP)	49
Resultados negativos	2 (FN)	45 (TN)	47
Total	46	50	96

Foi estabelecido um nível de significância de $\alpha = 5\%$. Nos experimentos realizados, $n_G = 46$, $n_H = 60$, $p_G = 0.957$ and $p_H = 0.667$. Foi obtido $\rho = 0.792$ e $Z = 3.647$. Com $\alpha = 5\%$ tem-se $Z_{\frac{\alpha}{2}} = 1.96$, o que corresponde à região de rejeição da hipótese H_0. No entanto, o valor P calculado foi de $0,0241 \leq 0,05$ e o poder do teste é 0,4225 (42,25%), portanto H_0 foi rejeitado.

Na Figura 4.3 é possível ver os cinco melhores resultados obtidos no primeiro experimento usando filtros de ICA com a base mini-MIAS.

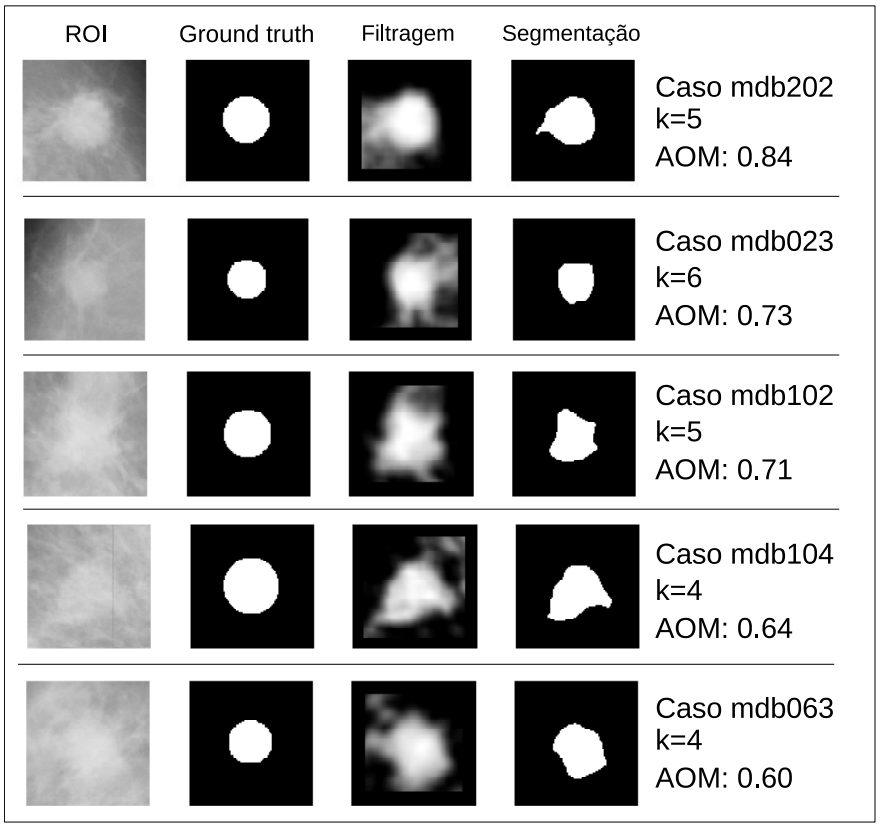

Figura 4.3: *5 melhores resultados obtidos no primeiro experimento usando filtros de ICA: base mini-MIAS*

Na Figura 4.4 é possível ver os cinco piores resultados obtidos no primeiro experimento usando filtros de ICA com a base mini-MIAS.

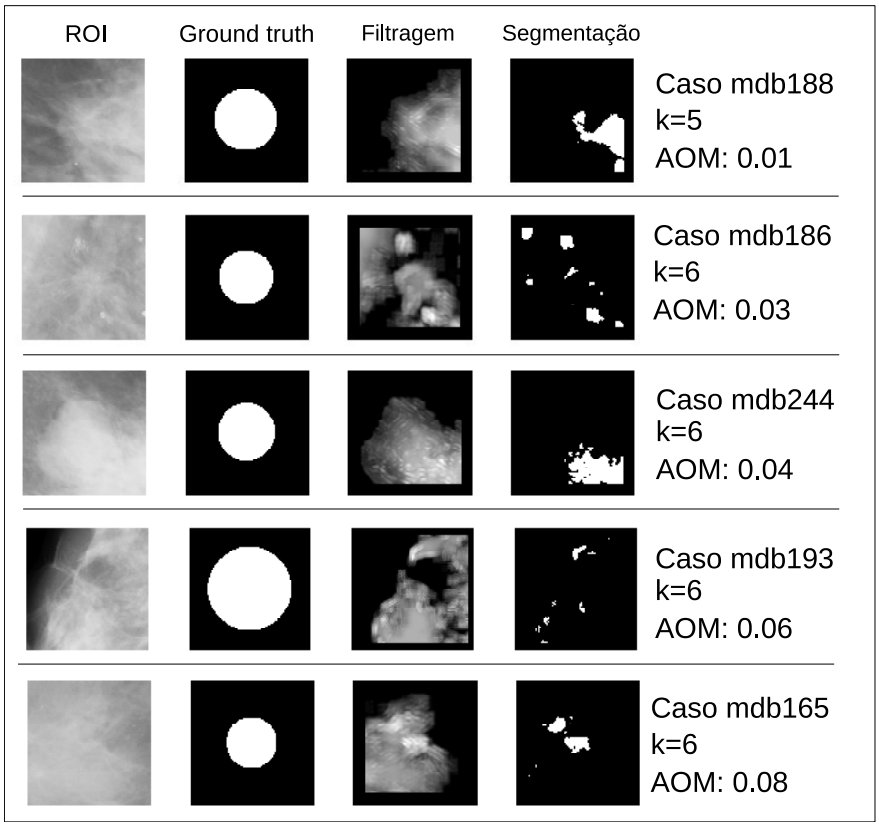

Figura 4.4: *5 piores resultados obtidos no primeiro experimento usando filtros de ICA: base mini-MIAS*

Na Figura 4.5 é possível ver os cinco melhores resultados obtidos no primeiro experimento usando filtros de ICA com a base DDSM.

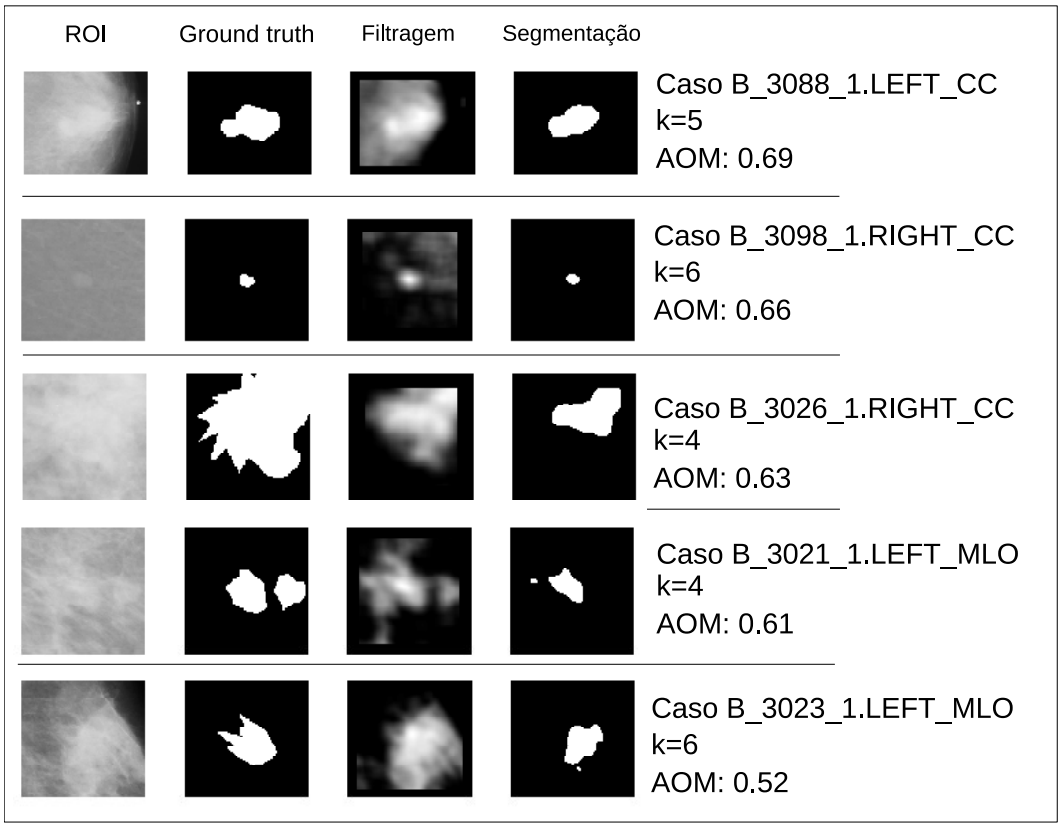

Figura 4.5: *5 melhores resultados obtidos no primeiro experimento usando filtros de ICA: base DDSM*

Na Figura 4.6 é possível ver os cinco piores resultados obtidos no primeiro experimento usando filtros de ICA com a base DDSM.

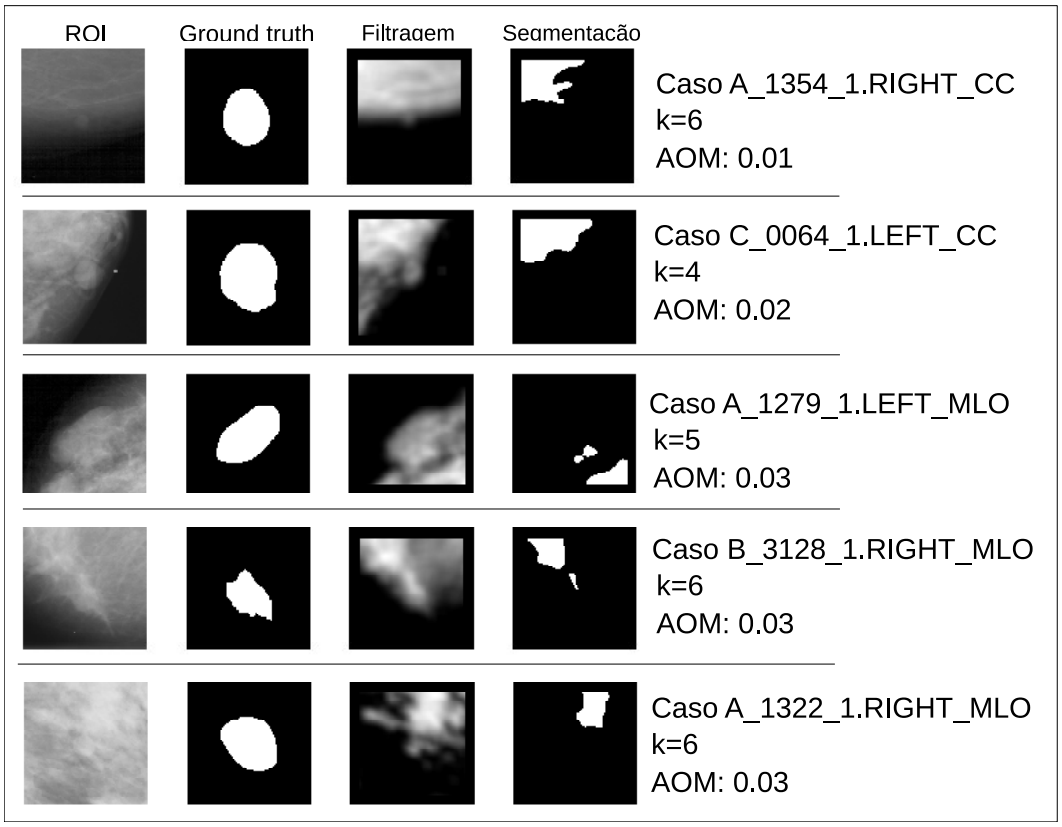

Figura 4.6: *5 piores resultados obtidos no primeiro experimento usando filtros de ICA: base DDSM*

Capítulo 5

Discussões

Após a obtenção dos resultados no primeiro experimento, observou-se que a taxa de acerto em mamas densas usando filtros de ICA foi maior que a taxa de acerto usando filtros de PCA. Isso ocorreu ao usar todas as funções base de ICA e ao selecionar as funções base com maior variância e maior curtose, conforme mostrado nas Tabelas 4.1, 4.2 e 4.3. Na Figura 5.1, é possível observar estes resultados, bem como a taxa de acerto quando aumentamos gradualmente o número de grupos de 4 para 6. Observou-se que quando foram usadas todas as funções base, os resultados são ligeiramente maiores do que quando se usam as funções base selecionadas com variância ou curtose. No entanto, a taxa de sucesso diminui à medida que foi aumentado o número de grupos.

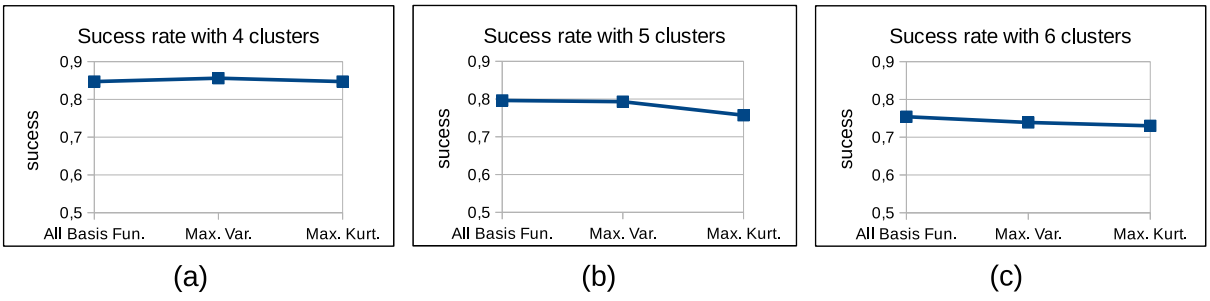

Figura 5.1: *Resultados da detecção de lesões usando filtros de ICA para 4, 5 e 6 grupos e usando (a) todas as funções base de ICA, (b) funções base de ICA de maior variação e (c) funções base de ICA de maior curtose.*

Além disso, como pode ser visto na Figura 5.2, os filtros de ICA apresentaram maiores taxas de detecção do que os filtros de PCA, independentemente de terem sido utilizadas todas as funções base ou apenas as funções base selecionadas. Tal como acontece com os filtros de ICA, ao aumentar o número de grupos, a taxa de acerto dos filtros de PCA também diminuiu.

Como se pode ver nos resultados, a taxa de acerto está relacionada ao número de

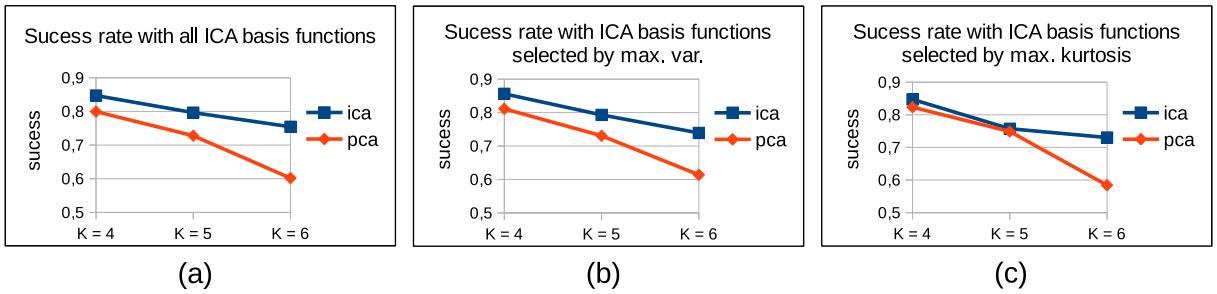

Figura 5.2: *Resultados da detecção de lesões usando filtros de ICA para 4, 5 e 6 grupos e usando (a) todas as funções base de ICA, (b) funções base de ICA de maior variação e (c) funções base de ICA de maior curtose.*

grupos definidos na segmentação. Acreditamos que o número de pixels contidos no grupos de maior energia tende a diminuir, uma vez que o conjunto total de pixels filtrados terá mais grupos. Uma vez que a métrica Jaccard é uma relação entre os pixels detectados pelos filtros e os pixels considerados como *ground truth*, essa taxa irá diminuir à medida que a acurácia diminui.

Na Figura 5.3, pode-se observar o resultado do diagnóstico sugerido fornecido pelo método proposto, após os filtros selecionados. Pode-se ver na Figura 5.3(a), Figura 5.3(b), Figura 5.3(c) e Figura 5.3(d) o resultado do processamento de imagem de duas mamografias que possuem algum tipo de massa, de acordo com os registros da base de dados Mini-MIAS. Podemos ver que o método destaca alguns verdadeiros positivos, no entanto, é sensível a falsos positivos.

Na Figura 5.4, pode-se observar alguns exemplos de lesões detectadas nos testes realizados com as bases de dados Mini-MIAS e DDSM. Pode-se observar a ROI contendo a lesão, o *ground truth* definida pelo especialista e o resultado segmentação variando os grupos na segmentação de 4 a 6 usando filtros de ICA, de maior variância e usando os filtros de PCA. Como se pode ver, nos casos A_1271_1.LEFT_MLO, A_1304_1.RIGHT_CC e A_1309_1.RIGHT_MLO, à medida que o número de grupos aumenta, a acurácia da detecção de lesões aumenta, enquanto os falsos positivos diminuem. Além disso, pode-se ver nos casos mdb001, mdb063, mdb081 e mdb104, que os filtros de ICA são mais precisos do que os filtros de PCA, pois delimitam mais claramente a borda da lesão. Assim, acredita-se que os resultados são mais favoráveis para os filtros de ICA tanto no caso de detecção quanto na definição do contorno de lesões em mamas densas.

O teste estatístico aplicado no segundo experimento permitiu verificar que, no nível de 5% de significância, há diferença estatística na acurácia obtida ao detectar lesões em mamas densas em comparação com a detecção de lesões em mamas não densas. Assim, pode-se perceber que a densidade da mama é um fator que altera a acurácia na detecção de lesões em mamas mais densas, quando se utilizam filtros de ICA.

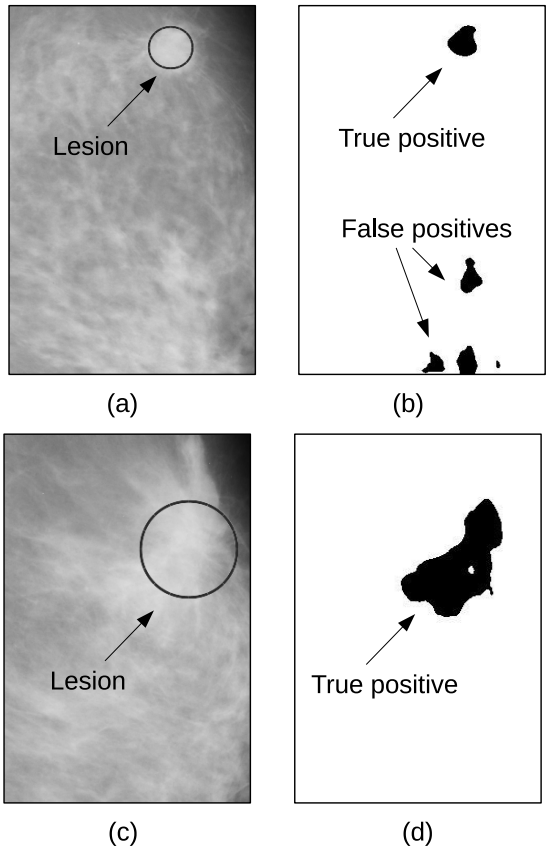

Figura 5.3: *Resultado final do processamento realizado em duas mamografias que possuem algum tipo de massa. (a) e (c) imagem analisada com indicação de massa de acordo com os registros da base de dados Mini-MIAS. (b) e (d) resultado da filtragem realizada nas imagens destacando regiões suspeitas.*

García-Manso et al. (2013) mostraram que o desempenho de seu método é afetado pela densidade mamária. Nossos resultados mostraram que a tarefa de segmentação baseada em filtros de ICA tem diferenças significativas quando aplicadas a mamas densas e não densas.

No entanto, percebeu-se que García-Manso et al. (2013) se concentraram na tarefa de classificação de lesões, usando ICA em todo o banco de dados DDSM. Neste estudo, o foco foi a tarefa de segmentação de massas, sem realizar classificação, usando o banco de dados Mini-MIAS e DDSM. Embora a comparação seja difícil neste caso, conclui-se que a densidade da mama pode afetar a segmentação e a classificação das regiões de interesse, utilizando filtros de ICA.

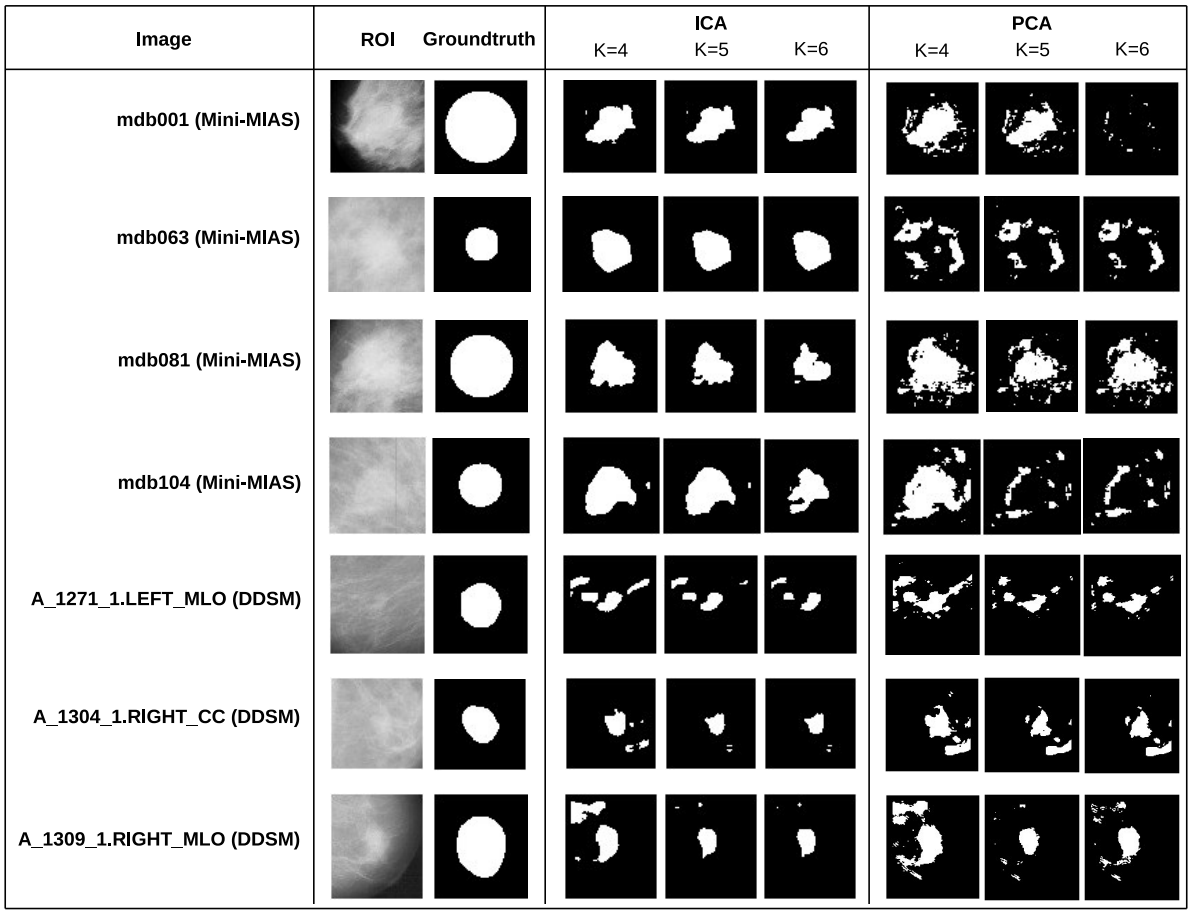

Figura 5.4: *Exemplos de lesões detectadas nos testes usando filtros de ICA de maior variância e usando filtros de PCA, aumentando o número de grupos na segmentação de 4 para 6.*

Capítulo 6

Conclusões e Atividades Futuras

Neste trabalho, foi apresentado um método para a detecção e segmentação de massas em imagens de mamas densas usando ICA. O método proposto objetiva auxiliar especialistas das áreas médica e radiológica, na tarefa de localização e classificação de massas em imagens de mamografia. Apesar da mamografia ser um dos exames mais comuns, sua sensibilidade pode variar de 46% a 88%, dependendo de fatores tais como tamanho e localização da lesão, densidade da mama, qualidade da mamografia e habilidade do especialista.

A densidade da mama, definida a partir da relação entre o tecido fibrograndular e o tecido adiposo presente na mama, pode dificultar a análise da mamografia. Isto ocorre em virtude dos tecidos fibrograndulares apresentarem densidade óptica e aparecerem mais claros na imagem. Como microcalcificações e massas também apresentam tonalidades mais claras na imagem de mamografia, muitas vezes se torna difícil para o especialista detectar lesões. Tal dificuldade foi a motivação para o desenvolvimento desta pesquisa.

A abordagem aqui utilizada foi baseada na extração de um conjunto de filtros obtidos de regiões de imagens não densas que contém massas para usá-los na detecção de massas em imagens de mama densas, cuja densidade é III ou IV de acordo com a escala BI-RADS. O método aqui proposto foi comparado com PCA.

Foram realizados dois experimentos usando as bases de dados Mini-MIAS e DDSM. Foram extraídas 36 ROIs da base Mini-MIAS e 83 ROIs da base DDSM, de imagens de mamografia de mamas não densas, para o processo de treinamento e obtenção dos filtros de ICA e PCA. O primeiro experimento teve por objetivo testar quais técnicas poderiam detectar e segmentar melhor a lesão, e todos os experimentos mostraram que os filtros de ICA apresentam melhor desempenho na segmentação de lesões em 139 ROIs extraídas das duas bases. No segundo experimento, o objetivo foi obter a acurácia dos filtros de

ICA na detecção da lesão em mamas densas e não densas. Foram avaliadas 120 ROIs e foi atingida uma acurácia de 79,17%, com sensibilidade de 66,67% e especificidade de 91,67%.

Trabalhos anteriores (BREM. et al., 2005; OLIVER et al., 2010) afirmaram que as técnicas CAD não são impactadas pela densidade mamária. Contrariamente a essa conclusão, mostramos que a detecção de massas utilizando o método baseado em filtros de ICA tem diferenças significativas quando aplicado em mamas densas e mamas não densas.

6.1 Trabalhos Futuros

Como proposta de trabalhos futuros, sugere-se investigar outros métodos de seleção das melhores funções base para compor um conjunto de filtros com o mínimo de redundância. Além disso, propõe-se aplicar uma técnica de redução de falsos negativos no diagnóstico, bem como classificar o tipo de lesão (benigno ou maligno) usando técnincas de aprendizagem de máquina.

6.2 Trabalhos Aceitos para Publicação

Silva, L. C. O.; Lopes, M. V.; Barros, A. K. Detecting masses in dense breast using independent component analysis. *Artificial Intelligence in Medicine*, 2017.

Referências Bibliográficas

ABDEL-QADER, I.; ABU-AMARA, F. A computer-aided diagnosis system for breast cancer using independent component analysis and fuzzy classifier. *Modelling and Simulation in Engineering*, v. 2008, p. 1–9, 2008.

ABU-AMARA, F.; ABDEL-QADER, I. Detection of breast cancer using independent component analysis. *Electro/Information Technology*, p. 428–431, 2007.

ACR. *Breast Imaging Reporting and Data System. BI-RADS*. 3. ed. [S.l.]: ACR, 1998.

ACS. *American Cancer Society*. 2017. <https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/types-of-breast-cancer/>. [Online; accessed 22-Jul-2017].

ANGELO, M. F. *Sistema de Processamento de Imagens Mamográficas e Auxílio ao Diagnóstico via-Internet*. Tese (Doutorado) — Escola de Engenharia de São Carlos da Universidade de São Paulo, 2007.

ARMITAGE, P. *Statistical methods in medical research*. 4. ed. [S.l.]: Wiley-Blackwell, 2001.

BARTLETT, M. S.; MOVELLAN, J. R.; SEJNOWSKI, T. J. Face recognition by independent component analysis. *IEEE transactions on neural networks*, v. 13, p. 1450–1464, 2002.

BERG, W. A. et al. Breast imaging reporting and data system. *American Journal of Roentgenology*, v. 6, 2000.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer-Verlag New York, Inc., 2006.

BOYD, N. F. et al. Mammographic density and the risk and detection of breast cancer. *New England Journal of Medicine*, v. 356, n. 3, p. 227–236, 2007.

BREM., R. F. et al. Impact of breast density on computer-aided detection for breast cancer. *AJR. American journal of roentgenology*, v. 184, 2005.

CAMPOS, L. F. A. et al. Segmentation and classification of breast cancer using independent component analysis, texture features and neural networks. In: Sociedade Brasileira de Computação. *WIM 2011 XI Workshop de Informática Médica*. [S.l.], 2011. p. 1764–1773.

CAVALCANTE, A. et al. Segmentation of natural and man-made structures by independent component analysis. *Lecture Notes in Computer Science*, v. 5441, p. 483–490, 2009.

CHIEN, J.-T.; CHEN, B.-C. A new independent component analysis for speech recognition and separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, v. 14, n. 4, p. 1245–1254, July 2006.

CHOI, S. et al. Blind source separation and independent component analysis: A review. *Neural Information Processing - Letters and Reviews*, v. 6, 2005.

CHRISTOYIANNI, I. et al. Computer aided diagnosis of breast cancer in digitized mammograms. *Computerized Medical Imaging and Graphics*, v. 26, n. 5, p. 309–319, 2002.

COSTA, D. D. *Processamento e Análise de Sinais Mamográficos na Detecção do Câncer de Mama: diagnóstico auxiliado por computador (CAD)*. Tese (Doutorado) — Universidade Federal do Maranhão, 2012.

GALLARDO-CABALLERO, R. et al. Independent component analysis to detect clustered microcalcification breast cancers. *The Scientific World Journal*, v. 2012, 2012.

GARCÍA-MANSO, A. et al. Study of the effect of breast tissue density on detection of masses in mammograms. *Computational and Mathematical Methods in Medicine*, v. 2013, 2013.

GONZALEZ, R. C.; WOODS, R. E.; EDDINS, S. L. *Digital Image Processing Using MATLAB*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2003.

HAIR JR., J. F. et al. *Análise Multivariada de Dados*. 6. ed. [S.l.]: Bookman, 2009.

HEATH, M. et al. The digital database for screening mammography. In: *Proceedings of the Fifth International Workshop on Digital Mammography*. [S.l.: s.n.], 2001. p. 212–218.

HO, W.; LAM, P. Clinical performance of computer-assisted detection (cad) system in detecting carcinoma in breasts of different densities. *Clinical Radiology*, v. 58, n. 2, p. 133–136, 2003.

HOYER, P. O.; HYVARINEN, A. Independent component analysis applied to feature extraction from colour and stereo images. In: *Network Computation in Neural Systems*. [S.l.: s.n.], 2000. p. 191–210.

HYVARINEN, A. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, IEEE, v. 10, n. 3, p. 626–634, 1999.

HYVARINEN, A. *The FastICA MATLAB package*. 2017. <http://www.cis.hut.fi/projects/ica/fastica/>. [Online; accessed 04-May-2017].

HYVARINEN, A.; KARHUNEN, J.; OJA, E. *Independent component analysis*. [S.l.]: John Wiley & Sons, 2004. v. 46.

HYVARINEN, A.; OJA, E. Independent component analysis: algorithms and applications. *Neural Networks*, v. 13, 2000.

INCA. *A mama: tratamento compreensivo das doenças benignas e malignas*. São Paulo: Manole, 1994.

INCA. *Mamografia: da prática ao controle*. Rio de Janeiro: INCA, 2007.

JAIN, A. K.; F., F. Unsupervised texture segmentation using gabor filters. *Pattern Recognition*, v. 24, p. 1167–1186, 1991.

JAMES, C. J.; HESSE, C. W. Independent component analysis for biomedical signals. *Physiological Measurement*, v. 26, n. 1, p. R15–R39, 2005.

JENSSEN, R.; ELTOFT, T. Independent component analysis for texture segmentation. *Pattern Recognition*, v. 36, n. 10, p. 2301–2315, 2003.

MOK, P.; LAM, K.; NG, H. An ica design of intraday stock prediction models with automatic variable selection. In: *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*. [S.l.: s.n.], 2004. v. 3, p. 2135–2140.

MOREIRA, I. C. et al. Inbreast: Toward a full-field digital mammographic database. *Academic Radiology*, v. 19, 2012.

OBENAUER, S. et al. Impact of breast density on computer-aided detection in full-field digital mammography. *Journal of Digital Imaging*, v. 19, n. 3, p. 258–263, 2006.

OLIVER, A. et al. Impact of breast density on computer-aided detection for breast cancer. *Academic radiology*, v. 17, 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, p. 607–609, 1996.

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

POLAK, M.; ZHANG, H.; PI, M. An evaluation metric for image segmentation of multiple objects. *Image and Vision Computing*, v. 27, p. 1223–1227, 2009.

RIBEIRO, P. B. et al. Automatic segmentation of breast masses using enhanced ica mixture model. *Neurocomputing*, Elsevier, v. 120, p. 61–71, 2013.

SILVA, W.; MENOTTI, D. Classification of mammograms by the breast composition. In: *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*. [S.l.: s.n.], 2012. p. 1.

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

SUCKLING, J. et al. The mammographic image analysis society digital mammogram database exerpta medica. In: *International Congress Series*. [S.l.: s.n.], 1994. v. 1069, p. 375–378.

TANG, J. et al. Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances. *Trans. Info. Tech. Biomed.*, v. 13, n. 2, p. 236–251, 2009.

THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern recognition*. [S.l.]: Academic Press, 2009.

VOROBYOV, S.; CICHOCKI, A. Blind noise reduction for multisensory signals using ica and subspace filtering, with application to eeg analysis. In: *Biological Cybernetics*. [S.l.: s.n.], 2002. p. 293–303.

WINSBERG, F. et al. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. *Radiology*, v. 89, n. 2, p. 211–215, 1967.

YANG, S. K. et al. Screening mammography-detected cancers: Sensitivity of a computer-aided detection system applied to full-field digital mammograms. *Radiology*, v. 244, n. 1, p. 104–111, 2007.