UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

LIZIANNE PRISCILA MARQUES SOUTO

MINERAÇÃO DE IMAGENS PARA A CLASSIFICAÇÃO DE TUMORES DE MAMA

MOSSORÓ - RN
2014

Lizianne Priscila Marques Souto

Mineração de Imagens para a Classificação de Tumores de Mama

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Computação - associação ampla entre a Universidade do Estado do Rio Grande do Norte e a Universidade Federal Rural do Semi-Árido, para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Marcelino Pereira dos Santos Silva

MOSSORÓ - RN
2014

Lizianne Priscila Marques Souto

Mineração de Imagens para a Classificação de Tumores de Mama

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Computação para a obtenção do título de Mestre em Ciência da Computação.

APROVADA EM: ___ / ___ / _____.

BANCA EXAMINADORA

_____

Prof. Dr. Marcelino Pereira dos Santos Silva - UERN
Presidente

_____

Prof. Dr. Carlos Heitor Pereira Liberalino - UERN
Primeiro Membro

_____

Prof. Dr. José Alfredo Ferreira Costa - UFRN
Segundo Membro

Catalogação da Publicação na Fonte.
Universidade do Estado do Rio Grande do Norte.

Souto, Lizianne Priscila Marques Souto
Mineração de imagens para a classificação de tumores de mama. / Lizianne Priscila Marques Souto. – Mossoró, RN, 2013.

107 f.
Orientador: Prof. Dr. Marcelino Pereira dos Santos Silva

Dissertação (Mestrado) Universidade do Estado do Rio Grande do Norte. Universidade Federal Rural do Semi-Árido. Programa de Pós-Graduação em Ciência da Computação.

1. Processamento digital de imagens. 2. Câncer de mama - Diagnóstico - Auxiliado por computador. 3. Mineração de dados. 4. Mineração de imagens. I. Silva, Marcelino Pereira dos Santos. II.Universidade do Estado do Rio Grande do Norte. III. Universidade Federal Rural do Semi-Árido. IV. Título.

UERN/BC CDD 004

Bibliotecária: Jocelania Marinho Maia de Oliveira – CRB 15 319

Ao que tenho de mais importante em minha vida: meus pais, Luciene e Hélio; e meus irmãos, Lizandro e Lucas.

AGRADECIMENTOS

A Deus, O único digno de toda honra e glória, pelo dom da vida e bençãos sem medidas.

Aos meus pais, Luciene e Hélio, pelo amor e apoio incondicional em todas as situações, sempre me instruindo no caminho certo e me apoiando em todas as decisões. Bem como aos meus irmãos, Lizandro e Lucas, meus melhores amigos e companheiros. A eles todo meu amor e gratidão.

Ao meu orientador Marcelino Pereira, o qual serei sempre grata pela confiança em mim depositada, pelo incentivo, apoio e ensinamentos que levarei para toda a vida. Um verdadeiro exemplo de ser humano, competência e profissionalismo a ser seguido.

A Thiago Kleyton, pela fundamental contribuição e comprometimento com este trabalho.

A Rosita, secretária do mestrado, a qual foi meu primeiro contato em Mossoró, sempre muito prestativa e amiga.

Aos colegas de mestrado e integrantes do Laboratório de Engenharia de Software - LES, em especial a Irlan, Rodrigo, Natan e Ciro, pelo companheirismo e trocas de conhecimentos.

Aos amigos que fiz em Mossoró, Nina, Kleber, Iracema, Patrícia e as trigêmeas (Carolina, Gabriela e Lorena), os quais compartilhei momentos de alegria e que por muitas vezes supriram a ausência da minha família.

Ao Centro de Oncologia e Hematologia de Mossoró - COHM, Marilyn Christine e a Karla Haryanna pelo apoio técnico-científico a esta pesquisa.

A UERN e a UFERSA pela oportunidade de aperfeiçoamento acadêmico e infraestrutura fornecida, bem como à CAPES e FAPERN pelo apoio financeiro.

Por fim, gostaria de agradecer aos meus amigos, familiares e professores que contribuíram de maneira direta ou indireta para a minha formação e realização deste trabalho.

Mas os que esperam no Senhor renovarão as forças, subirão com asas como águias; correrão, e não se cansarão; caminharão, e não se fatigarão.

ISAÍAS 40:31

RESUMO

Apesar da mamografia ser considerada o melhor método de detecção precoce do câncer de mama, por motivos como limitações próprias do especialista, cerca de 10% a 30% das lesões mamárias não são identificadas. Uma solução para esse problema é a utilização de sistemas de Diagnóstico Auxiliado por Computador (CAD) que permitem uma melhor análise de imagens, aumentando em até 15% a sensibilidade no rastreamento de mamografia. Neste trabalho é proposta uma metodologia e um sistema CAD para auxílio ao diagnóstico de câncer de mama. A abordagem inclui a extração de atributos geométricos de tumores em mamografias e o uso de algoritmos de mineração de dados para a classificação de tais tumores como benignos ou malignos. Como produto desta dissertação tem-se o software MAMOCAD, o qual mostrou-se eficaz para a diminuição de casos de falso positivo e falso negativo, tornando o processo de tomada de decisão do especialista mais eficiente e preciso.

Palavras-chave: Câncer de mama, Mineração de dados, Mineração de imagens, Diagnóstico auxiliado por computador, Processamento digital de imagens.

ABSTRACT

Although mammography is considered the best method for early detection of breast cancer, for reasons such as limitations of specialist, about 10% to 30% of breast lesions are not identified. One solution to this problem is the use of Computer-Aided Diagnosis (CAD) that provides a better image analysis, increasing up to 15% the sensitivity in mammographic screening. In this work a methodology and a CAD system for the diagnosis of breast cancer is proposed. The approach includes the extraction of geometric features from breast tumors in mammograms and the use of data mining algorithms for classification of such tumors as benign or malignant. The product of this dissertation is the MAMOCAD software which is effective for the reduction of false positive and false negative cases, becoming the medical decision process more efficient and accurate.

Keywords: Breast cancer, Data mining, Image mining, Computer-aided diagnosis, Digital image processing.

LISTA DE SIGLAS

AC	Acurácia
ACR	American College of Radiology
ACS	Sociedade Americana de Câncer
AD	Árvore de Decisão
Adaline	Adaptive Linear Neuron
AM	Aprendizagem de Máquina
API	Interface de Programação de Aplicações
BI-RADS	Breast Image Reporting and Data System
CAD	Computer-Aided Diagnosis
CADe	Computer-Aided Detection
CADx	Computer-Aided Diagnosis
CART	Classification and Regression Trees
CBR	Colégio Brasileiro de Radiologia
CC	Crânio-Caudal
CDF	Cumulative Distribution Function
DCBD	Descoberta de Conhecimento em Banco de Dados
DDSM	Digital Database for Screening Mammography
DICOM	Digital Imaging and Communications in Medicine
E	Especificidade
FB	Crânio-Caudal de Baixo

FN	Falso Negativo
FP	Falso Positivo
GI	Ganho de Informação
ID3	Iterative Dichotomiser
IDE	Integrated Development Environment
INCA	Instituto Nacional do Câncer
ISO	Ínfero Superior Oblíqua
JAI	Java Advanced Imaging
KDD	Knowledge Discovery in Databases
K-NN	K-Nearest Neighbour
LM	Látero Medial
LMO	Lateral Médio Oblíqua
MAMOCAD	Computer-Aided Diagnosis in Mammography
MD	Mineração de Dados
MIAS	Mammographic Image Analysis Society
ML	Médio Lateral
MLO	Médio Lateral Oblíqua
MLP	Multilayer Perceptron
MM	Morfologia Matemática
MVC	Model View Control
OMS	Organização Mundial da Saúde

PAAF	Punção Aspirativa por Agulha Fina
PDI	Processamento Digital de Imagens
RG	Region Growing
RM	Ressonância Magnética
RNA	Rede Neural Artificial
ROI	Region of Interest
S	Sensibilidade
SOI	Súpero Inferior Oblíqua
SOM	Self-Organizing Maps
SVM	Support Vector Machine
TRH	Terapia de Reposição Hormonal
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
VPN	Valor Preditivo Negativo
VPP	Valor Preditivo Positivo
WEKA	Waikato Enviroment for Knowledge Analysis

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO 1

INTRODUÇÃO

O aumento no surgimento de novos casos de câncer de mama tem repercutido como um dos graves problemas de saúde pública no mundo. Países em desenvolvimento, como o Brasil, são os mais afetados com essa doença devido à falta de informação e precarização de parte dos serviços de saúde enfrentados pela população. O câncer de mama também é uma das principais causas de mortes em mulheres com idade entre 39 e 58 anos. No ano de 2010 cerca de 12.705 mulheres tiveram mortes ocasionadas pelo câncer de mama no Brasil (INCA, 2012). A Organização Mundial da Saúde (OMS) projeta para o ano de 2030 uma média de 27 milhões de novos casos de câncer de mama no mundo, o que resultará em 17 milhões de mortes.

O combate a esta doença se dá por meio da realização periódica do auto-exame das mamas, do exame clínico realizado pelo profissional de saúde e de exames por imagens, como a mamografia, pois com a descoberta precoce existe a possibilidade da realização de tratamentos que aumentam as chances de cura do paciente. A mamografia, comparada aos demais exames apresentados, é a forma mais eficaz para a detecção de tumores em fase inicial, a partir de 1 mm, enquanto no auto-exame só é possível perceber tumores a partir de 1,5 cm (LEITE et al., 2011). A mamografia apresenta até 88% de sensibilidade e até 99% de especificidade (INCA, 2002). A sensibilidade é a capacidade de detectar o verdadeiro positivo, ou seja, caso em que existe a doença. Já a especificidade está relacionada à detecção do verdadeiro negativo, em outras palavras significa a detecção correta de casos livres de câncer.

Além da mamografia existem outras modalidades de imagens médicas para o diagnóstico do câncer de mama, dentre elas estão: ultrassonografia, tomografia computadorizada, ressonância magnética e termografia. Apesar dessa grande diversidade de tipos de imagens e do sistema de padronização de laudos BI-RADS, dependendo do caso, os especialistas podem encontrar dificuldade na identificação e interpretação das lesões nas imagens, visto que elas podem se apresentar em diferentes posições na mama e, dependendo do tipo, a morfologia e tamanho variam. Estima-se que a sensibilidade dos radiologistas no rastreio do câncer de mama varia entre 65% e 75% (SKAANE; ENGEDAL; SKJENNALD, 1997) e é influenciada por limitações pró-

prias do especialista como cansaço, indisposição, entre outros. Então, como consequência da má interpretação dos exames por imagens e da insegurança no diagnóstico, tem-se o aumento do número de realização de biópsias desnecessárias e de tratamentos inadequados.

Uma alternativa viável para auxiliar os radiologistas na leitura de imagens e diagnóstico precoce do câncer de mama é a utilização de esquemas de Diagnóstico Auxiliado por Computador (*Computer-Aided Diagnosis (CAD)*). Os sistemas CAD não substituem o especialista, pelo contrário, são úteis para diminuir a incerteza no diagnóstico fornecendo uma "segunda opinião" do caso. De modo geral, os sistemas CAD podem dispor de técnicas para melhorar a qualidade da imagem e, consequentemente, a visualização e localização de lesões suspeitas, extrair características a partir de imagens e classificar os achados de acordo com sua probabilidade de malignidade.

Por outro lado, a mineração de dados é uma das etapas do processo de Descoberta de Conhecimento em Banco de Dados (DCBD), que tem como objetivo a aplicação de técnicas e métodos computacionais para extrair informações úteis, válidas e novas a partir de um grande conjunto de dados para a tomada de decisão. A mineração pode ser aplicada tanto em repositórios de dados convencionais como em repositórios de dados não convencionais, como é o caso das imagens (SIMOFF; DJERABA; ZAIANE, 2002). Quando aplicada a imagens, "trata-se da extração de conhecimento implícito, relacionamento de imagens e outros padrões não explicitamente armazenados em bases de dados de imagens" (ZHANG; HSU; LEE, 2002).

Diante da relevância do contexto exposto, nesta dissertação propõe-se responder às seguintes questões: a partir de imagens de mamografias, quais parâmetros avaliar para caracterizar tumores de mama e diferenciá-los como malignos ou benignos? Atributos geométricos caracterizam adequadamente tumores benignos e malignos em mamografias?

A partir destes questionamentos, neste trabalho é levantada a hipótese de que a partir de conceitos de Processamento Digital de Imagens (PDI) e Mineração de Imagens é possível desenvolver uma metodologia de extração de atributos a partir de regiões de interesse (do inglês *Region of Interest (ROI)*) em imagens de mamografia, os quais submetidos ao processo de mineração são classificadas de acordo com o tipo de lesão apresentada.

1.1 OBJETIVO GERAL

Neste trabalho é proposto o desenvolvimento de uma metodologia e um CAD mamográfico para auxiliar especialistas da área médica na classificação de tumores de mama como benignos ou malignos de maneira eficaz e eficiente.

1.2 OBJETIVOS ESPECÍFICOS

Como objetivos específicos deste trabalho podemos destacar:

- Investigar aspectos relevantes da definição e análise de morfologia (geometria) de regiões de interesse (tumores) em oncologia mamária;
- Implementar operadores de pré-processamento que realcem tumores em imagens mamográficas;
- Implementar algoritmos que segmentem satisfatoriamente imagens de mamografias;
- Implementar algoritmos que extraiam atributos geométricos de regiões de interesse obtidos a partir da segmentação de mamografias;
- Implementar algoritmos de mineração de imagens através dos atributos extraídos.

1.3 ORGANIZAÇÃO DA DISSERTAÇÃO

Esta dissertação encontra-se organizada em 6 capítulos, a saber: no Capítulo 2 é discutido o câncer de mama, imageamento mamográfico, classificação BI-RADS e sistemas CAD; no Capítulo 3 são apresentados conceitos sobre o processo de DCBD, mineração de dados e mineração de imagens. Também encontram-se descritas as etapas e técnicas de PDI, bem como uma revisão bibliográfica dos trabalhos relacionados; no Capítulo 4 é realizada uma descrição da metodologia e do sistema proposto (MAMOCAD); no Capítulo 5 encontra-se descrito o repositório de imagens BCDR, assim como o detalhamento dos experimentos realizados no estudo de caso e dos resultados obtidos; por fim, no Capítulo 6 são apresentadas as principais conclusões deste trabalho e os trabalhos futuros.

CAPÍTULO 2

CÂNCER DE MAMA E IMAGEAMENTO MAMOGRÁFICO

Este capítulo tem como objetivo apresentar aspectos relacionados ao câncer de mama e os fatores que contribuem para o aparecimento da doença, enfatizando a importância do diagnóstico precoce para aumentar as chances de cura do paciente. Também descreve-se sobre a anatomia da mama e faz-se referência às modalidades de imagens médicas que podem ser utilizadas para melhor avaliá-la. Por fim, discorre-se sobre a utilização de sistemas de diagnóstico auxiliado por computador e medidas de desempenho de tais sistemas.

2.1 CÂNCER DE MAMA

Câncer é o nome dado às doenças que apresentam em comum o crescimento desordenado de células que atingem tecidos e órgãos (INCA, 2014). Célula é a unidade básica do organismo, ela cresce e divide-se para a produção de mais células de acordo com a necessidade do organismo para manter-se saudável. Quando esse processo ocorre de forma rápida, desordenada e desnecessária, forma-se uma massa de tecido chamada de neoplasia ou tumor.

Os tumores podem ser não cancerígenos (benignos) ou cancerígenos (malignos). Tumores malignos tendem a penetrar e destruir tecidos saudáveis do corpo. Em casos mais avançados da doença pode ocorrer a metástase, fenômeno em que as células malignas se espalham pelos vasos sanguíneos para outras partes do corpo atingindo órgãos vitais como fígado, pulmão e cérebro, por exemplo.

As causas da doença podem estar relacionadas a fatores externos ou internos ao organismo, como por exemplo: condições ambientais (água, terra e ar), hábitos alimentares, estilo de vida, predisposição genética, entre outros. Com isso, os hábitos e o estilo de vida adotados por cada pessoa podem determinar diferentes tipos de câncer (INCA, 2014).

O câncer de mama, foco deste trabalho, é um câncer que atinge o tecido mamário. Os tipos mais frequentes se originam nos ductos lactíferos (canais que conduzem o leite até a papila) ou nos lóbulos mamários (responsáveis pela secreção de leite) (KIERSZENBAUM; TRES, 2012). Trata-se de uma das doenças que mais atinge as mulheres e que mais causa mortes no mundo.

Países em desenvolvimento como o Brasil são os mais afetados com esta doença seguido de um alto índice de mortalidade devido à sua descoberta já em estágio avançado, visto que a taxa de sobrevivência do paciente é proporcional ao estágio em que ela é detectada.

Na Figura 2.1 é possível observar as taxas de mortalidade de mulheres no Brasil, entre os anos de 1990 e 2011, de acordo com a localização do câncer (mama; traquéia, brônquios e pulmões; cólon e reto; colo do útero; e estômago). Nesta figura é possível constatar o aumento significativo da taxa de mortalidade causada pelo câncer de mama, o que o torna um dos principais problemas de saúde enfrentados pela população até a atualidade.

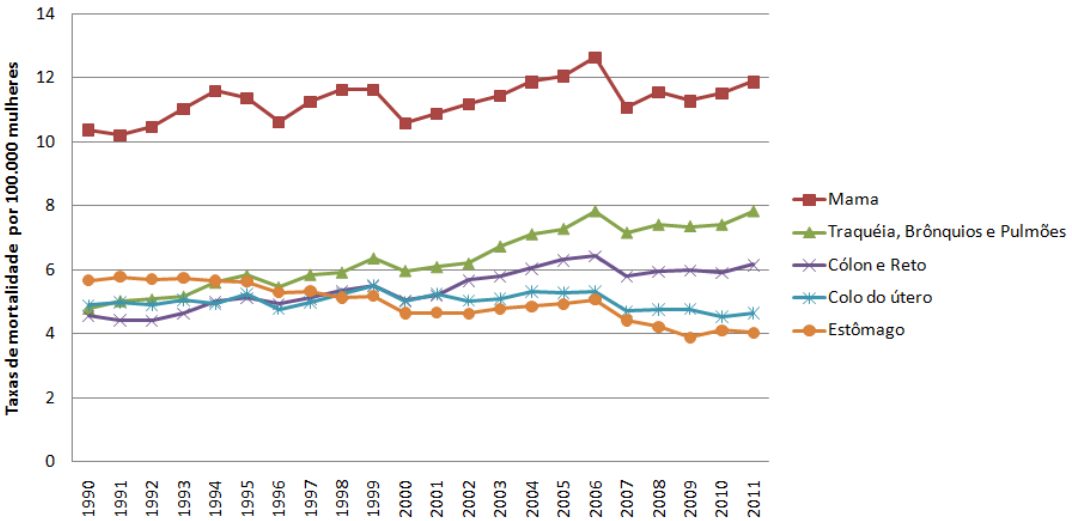

Figura 2.1: Taxas de mortalidade de 5 incidências de câncer mais frequentes
Fonte: (INCA, 2011)

Por se tratar de uma doença em que 10% dos casos é hereditário, o câncer de mama pode se desenvolver tanto em mulheres quanto em homens. Fatores como má alimentação, sobrepeso e exposição a hormônios femininos também podem contribuir para o aparecimento da doença. Pessoas que apresentam ou que já apresentaram câncer em uma das mamas (ipsilateral) podem facilmente obter câncer contralateral, ou seja, na outra mama (LIZARRAGA et al., 2013).

O câncer de mama em fase inicial geralmente não apresenta sintomas, por isso é importante o acompanhamento médico e a realização de exames periodicamente. De acordo com o crescimento do tumor, pode-se observar mudança no tamanho e no formato da mama ou do mamilo. Em casos mais avançados pode-se incluir sintomas como dor óssea, dor mamária ou desconforto, inchaço na axila (ao lado da mama com câncer), perda de peso, etc. Dessa forma

é recomendado, principalmente às mulheres a partir dos 35 anos, as seguintes medidas para diagnosticar o câncer de mama:

- Autoexame das mamas - método realizado pela própria mulher com o objetivo de encontrar nódulos ou anomalias nas mamas. Esse procedimento deve ser realizado cerca de 7 dias após o período da mestruação, visto que é comum o aparecimento de nódulos durante a mestruação e desaparecimento dos mesmos com o passar do tempo. Em casos em que a mulher não menstrua mais, deve ser realizado uma vez por mês a qualquer tempo. O autoexame pode ser realizado independentemente da faixa etária e não deve substituir o exame realizado pelo profissional de saúde;

- Exame clínico das mamas - realizado por um médico ou especialista da área de saúde. Consiste em apalpar as axilas e mamas para detectar lesões palpáveis, verificando seu tamanho, textura, mobilidade/fixação;

- Mamografia - rastreamento por mamografia é a forma mais segura para detectar a presença de lesões nas mamas, principalmente em estágio inicial. A mamografia não deve descartar o exame clínico realizado pelo médico, já que podem existir massas palpáveis que não são detectadas na mamografia.

Caso seja identificada alguma anomalia na mamografia o médico pode solicitar a realização de outro exame por imagem ou até mesmo uma biópsia da lesão para confirmar o diagnóstico. A biópsia é um exame invasivo solicitado pelo médico que pode ser cirúrgica ou não cirurgica (Punção Aspirativa por Agulha Fina (PAAF), por exemplo.). Apresenta como desvantagem o alto custo financeiro e desgaste emocional superior comparado a outros métodos de detecção como a mamografia e ultrassonografia.

Nota-se que a incerteza no diagnóstico tem contribuído para a realização de biópsias cirúrgicas desnecessárias. Cerca de 15 a 30% das massas encaminhadas para biópsia cirúrgica são realmente malignas (MOHAMED; KADAH, 2007), o que poderiam ser evitadas com a devida avaliação por meio de exames por imagens.

Após a confirmação do diagnóstico da doença é então realizado o tratamento baseando-se no tipo e estágio do câncer. Utilizando-se de medicamentos quimioterápicos, radioterapia, cirurgia

e/ou terapia hormonal, para destruir as células e o tecido canceroso, bloqueando hormônios que estimulam seu desenvolvimento.

2.2 ANATOMIA DA MAMA

As mamas são órgãos externos presentes tanto em mulheres quanto em homens. Nos homens elas não apresentam função e por isso permanecem rudimentares. Nas mulheres são desenvolvidas após a puberdade devido ao estímulo hormonal do estrógeno e progesterona.

Encontram-se situadas nas paredes ântero-laterais torácicas, da segunda à sexta costela, onde dois terços estão sobre a fáscia peitoral que cobre o músculo peitoral maior e um terço repousa sobre a fáscia que reveste o músculo serrátil anterior (MOORE; DALLEY, 2007).

As mamas são constituídas por três tipos de tecido: adiposo, fibroso (conjuntivo) e glândular epitelial (parênquima) (GRAY, 2005). Por sua vez, o parênquima mamário é composto por cerca de 15 a 20 lobos mamários e de seus respectivos ductos lactíferos que se ligam à papila, e tem como função a secreção de leite. A conexão ente os lobos é feita por meio do tecido conjuntivo e o intervalo entre eles é preenchido por tecido adiposo. Além disso, nas mamas são encontrados vasos sanguíneos, vasos linfáticos e fibras nervosas.

Na Figura 2.2 são representadas as principais estruturas encontradas na mama feminina e as respectivas descrições encontram-se nos itens abaixo de acordo com Moore e Dalley (2007):

- Ligamento suspensor (*Cooper*) - são ligamentos subcutâneos que provêm mobilidade e sustentação às mamas;
- Alvéolos - são unidades básicas do tecido glandular responsáveis pela produção de leite;
- Lóbulo mamário - conjunto de 10 a 100 alvéolos;
- Lobo mamário - conjunto de lóbulos mamários que se liga à papila através de um ducto lactífero;
- Ducto lactífero - dá origem a brotamentos que formam de 15 a 20 lóbulos de tecido glandular que constituem a glândula mamária. Os ductos conduzem o leite secretado até a papila mamária;
- Tecido glandular - conjunto de lobos e ductos;

- Seio lactífero - é uma porção dilatada do ducto que armazena o leite produzido;
- Aréola - estrutura central da mama onde se projeta a papila;
- Papila mamária - trata-se das protuberâncias cônicas ou cilíndricas nos centros das aréolas onde desembocam os ductos;
- Lóbulo de gordura - o restante da mama é preenchido por tecido adiposo/gorduroso, de acordo com a idade da mulher e suas condições físicas;
- Espaço retromamário - é um espaço ou plano potencial de tecido conectivo frouxo entre a mama e a fáscia peitoral, que permite à mama uma certa mobilidade na parede torácica;
- Fáscia peitoral - membrana que recobre os músculos.

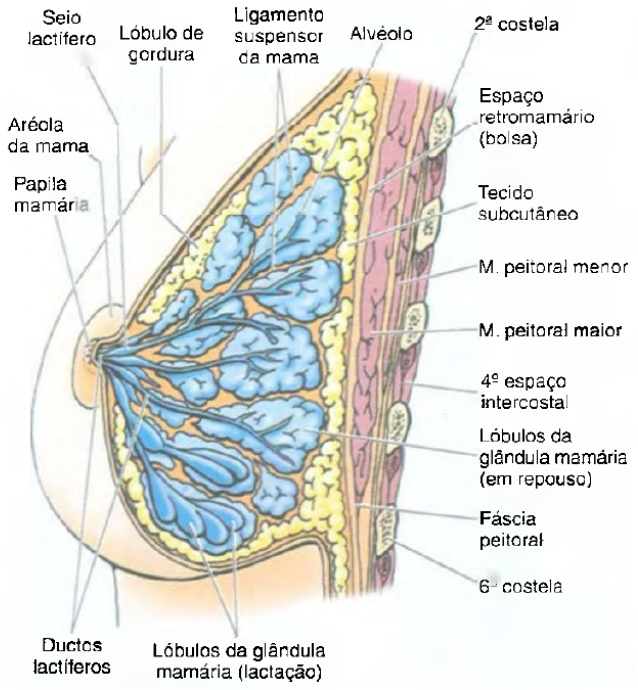

Figura 2.2: Estruturas da mama
Fonte: (MOORE; DALLEY, 2007)

Visando a facilitação da localização de possíveis alterações nas mamas, a superfície da mesma pode ser dividida em quadrantes: súpero lateral, súpero medial, ínfero lateral, ínfero medial (Figura 2.3). Também pode-se referir-se à localização por meio da analogia com o relógio (*clock face*), informando a posição do ponteiro para identificar a lesão (Figura 2.3).

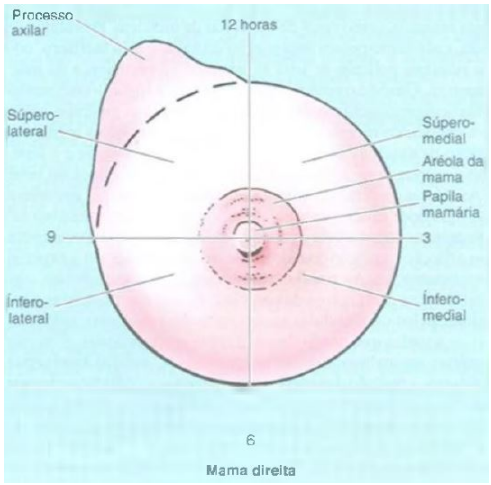

Figura 2.3: Quadrantes da mama
Fonte: (MOORE; DALLEY, 2007)

2.3 DENSIDADE MAMÁRIA

A densidade é descrita pela proporção da quantidade de tecido adiposo versus tecido glandular (denso) presente na mama (ERNSTOFF et al., 2006). Logo, quanto maior a quantidade de tecido adiposo, menor será a densidade mamária.

A densidade das mamas diferem particularmente de uma mulher para outra e está associada à idade, número de gestações, lactação, menopausa, Terapia de Reposição Hormonal (TRH), indice de massa corporal, entre outros fatores (MOUSA et al., 2013).

Evidências indicam que mulheres mais jovens, até 30 anos, apresentam mamas densas (até 90%) pelo fato de serem constituídas por uma grande quantidade de tecido glandular e com uma pequena quantidade de tecido adiposo (KOPANS, 2000). A medida que a mulher envelhece, entre 30 e 50 anos, ocorre a substituição do tecido glandular pelo tecido adiposo, resultando em mamas de densidade média (KOPANS, 2000).

Após os 50 anos a maior parte do tecido glandular se atrofia e gradativamente é substituído por tecido adiposo. Porém, não é possível afirmar uma correlação perfeita entre a faixa etária e a substituição do parênquima mamário por tecido adiposo (lipossubstituição), pois é possível que mulheres idosas apresentem mamas densas e mulheres jovens tenham mamas adiposas (FIGUEIRA et al., 2003). Principalmente pelo fato de que a realização de terapia de reposição hormonal para minimizar efeitos de menopausa em mulheres acima dos 50 anos têm mantido uma maior densidade mamária (ERNSTOFF et al., 2006).

Durante a gestação também ocorrem transformações que tornam as mamas mais densas, visto que, as glândulas aumentam de tamanho e um novo tecido glandular é formado para a produção do leite materno, o que diminui a quantidade de tecido adiposo (MOORE; DALLEY, 2007). Além disso, mulheres com apenas uma gestação tendem a apresentar mamas mais densas comparadas à mulheres com duas ou mais gestações (ERNSTOFF et al., 2006).

Além disso, estudos apontam que a densidade mamária é um forte indicador para determinar o risco de ocorrência do câncer de mama (VACEK; GELLER, 2004); (KERLIKOWSKE et al., 2007). Mulheres que apresentam tecido glandular predominante na mama têm cerca de 1,6 a 6 vezes mais chances de apresentarem a doença (CHEN et al., 2013). Um dos fatores desta associação deve-se à predisposição do tecido glandular em obscurecer a visibilidade e mascarar as lesões, reduzindo a sensibilidade e a especificidade do radiologista na leitura da mamografia (MOUSA et al., 2013).

Sendo assim, na literatura é possível encontrar métodos de classificação de padrões mamográficos de parênquimas associados ao risco de ocorrência do câncer de mama e à dificuldade de visualizar as lesões.

Dentre os quais, o pioneiro Wolfe (1976) baseou-se na proporção dos tecidos adiposo, glandular epitelial e fibroso, e ductos proeminentes observados na mamografia para propor uma das principais classificações de densidade mamária em quatro categorias:

- N1 - apresenta baixo risco de incidência de câncer. Parênquima é composto basicamente por gordura e pequena quantidade de displasia. Os ductos não são visíveis;
- P1 - apresenta risco intermediário de incidência de câncer. Parênquima composto principalmente de gordura com dutos proeminentes que ocupam até um quarto do seu volume;
- P2 - apresenta alto risco de incidência de câncer. Ductos proeminentes que ocupam mais que um quarto do volume da mama;
- DY - parênquima denso, obscurecendo um padrão ductal proeminente, apresentando alto risco de incidência de câncer.

Mais recentemente, *The American College of Radiology* propôs uma classificação similar à de Wolfe usada por radiologistas para avaliações mamográficas (VACEK; GELLER, 2004).

Essa classificação, discutida detalhadamente na Subseção 2.4.3, descreve as diferentes densidades da mama, de acordo com a quantidade de tecido fibroglandular (tecido mamário visto em mamografia), em quatro categorias BI-RADS (*Breast Imaging Reporting and Data System*) (RADIOLOGY, 1993):

- BI-RADS 1 (Gordurosa - lipossubstituída) - tecido adiposo predominante. Menos de 25% de tecido glandular;
- BI-RADS 2 (Parcialmente lipossubstituída) - tecido fibroglandular espalhado obscurecendo a lesão. Nível de tecido glandular entre 25% e 50%;
- BI-RADS 3 (Densa heterogênea) - mama heterogeneamente densa diminuindo a sensibilidade da mamografia. Nível de densidade entre 50% e 75%;
- BI-RADS 4 (Densa) - mama extremamente densa, acima de 75% de tecido glandular. Baixa sensibilidade mamográfica.

Na Figura 2.4 são apresentados exemplos de mamografias de acordo com os diferentes níveis de classificação BI-RADS. É possível observar que a densidade da mama aumenta de BI-RADS I (Figura 2.4 (a)) a BI-RADS IV (Figura 2.4 (d)).

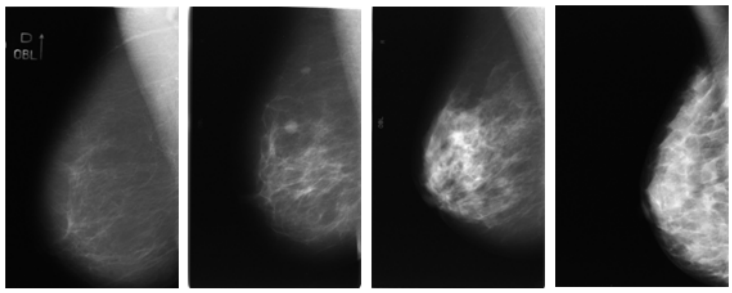

Figura 2.4: Classificação BI-RADS de acordo com a densidade mamária
Fonte: (BCDR, 2012)

Também influenciada por Wolfe, a modelagem Tabár baseada em correlações anatômicas-mamográficas divide as mamografias em 5 categorias (GRAM; FUNKHOUSER; TABAR, 1997):

- Categoria I - apresentam contornos enrugados e ligamentos de Cooper, unidades lobulares ductais terminais espalhadas e áreas transparentes de forma oval correspondente à formação de gorduras. Podendo evoluir para as categorias II ou III;

- Categoria II - representa substituíção completa de gordura;
- Categoria III - combinação de padrão ductal proeminente retroareolar devido elastose peridutal e interiorização de gordura;
- Categoria IV - densidades lineares e nodulares ao longo da mama;
- Categoria V - fibrose sem estrutura, homogênea, parecidas com vidro, e com contornos convexos.

As categorias de I a III estão relacionadas ao baixo risco de ocorrer o câncer de mama, já as categorias IV e V correspondem aos grupos de alto risco (GRAM; FUNKHOUSER; TABAR, 1997).

2.4 IMAGENS MÉDICAS

O diagnóstico de vários tipos de doenças, inclusive do câncer de mama, é suportado por exames de imagens. A aquisição de imagens médicas pode ser realizada por câmeras digitais, *scanners* e equipamentos (de raio-x, tomografia computadorizada, ultrassonografia e ressonância magnética) que utilizam o padrão *Digital Imaging and Communications in Medicine (DICOM)* (HENRIQUE NETO et al., 2003).

Existem várias modalidades de imagens médicas para o diagnóstico do câncer de mama. Cada método apresenta vantagens e desvantagens em sua utilização, podendo ser mais indicado em determinados casos. Dentre as modalidades de imagens médicas pode-se citar:

- Radiografia - é resultante da projeção de fótons de raios-x sobre o objeto em análise (GUY; FFYTCHE, 2005). A imagem de raio-x, neste caso a mamografia, apresenta diferentes tons de cinza, o que permite a visualização de massas, cistos, calcificações, entre outras anomalias. É uma técnica usual, de custo inferior comparado à tomografia ou à ressonância magnética e eficaz na descoberta do câncer de mama em estágio inicial. O exame deve ser feito com moderação, pois em excesso ocasiona complicações ao organismo humano devido à radiação ionizante provocada pelos raios-X;
- Ultrassonografia - é formada ao refletir ondas sonoras, com frequência de 1 a 10 MHZ, sobre a região em análise. Trata-se de um exame de baixo custo utilizado na detecção

de tumores na mama e complementar à mamografia. Comparada à mamografia, a ultrasonografia é mais eficaz em casos de pacientes com mamas densas, porém, ineficiente na detecção de microcalcificações. É útil principalmente na diferenciação entre lesões císticas e sólidas (PAULINELLI; CALAS; FREITAS JUNIOR, 2007), porém, ineficaz na distinção entre lesões sólidas benignas e malignas;

- Tomografia Computadorizada - técnica de aquisição de imagens baseada na radiografia convencional, porém, na tomografia os pacientes sofrem menos exposição à radiação. Nesta abordagem o tubo de raios-X gira 360 graus em torno do corpo obtendo diferentes pontos de vista da região investigada, em seguida, é realizada a reconstrução da imagem por meio do computador (GUY; FFYTCHE, 2005). A imagem obtida apresenta qualidade superior à mamografia, mas o fato de utilizar radiação X ainda a torna prejudicial à saúde;

- Ressonância Magnética (RM) - é utilizado um forte campo magnético e ondas de radiofrequência que emitem um sinal para cada amostra de núcleo de célula da região analisada, esses sinais são retornados formando a imagem. A RM proporciona imagens de maior contraste do que a técnica de raio-X e não utiliza radiação ionizante (GUY; FFYTCHE, 2005). Dentre todas as modalidades de imagens para a detecção do câncer de mama, a RM atinge a maior sensibilidade, se aproximando de 99% (KUHL, 2007). Apesar de ser a mais eficaz, também apresenta custos e riscos mais elevados que as demais modalidades. O fato de utilizar um campo magnético para a formação das imagens limita sua utilização, pois não é adequada para todos os pacientes, principalmente àqueles que possuem válvulas cardíacas, implantes auditivos e próteses dentárias. Também não é recomendado à mulheres grávidas ou pacientes alérgicos a determinados medicamentos, visto que, para a realização do exame é aplicado uma substância (contraste) intravenosa;

- Termografia - a imagem térmica é obtida por meio de uma câmera termográfica que capta a radiação infravermelha emitida pela pele (HART, 1991). Essa radiação é convertida em sinal elétrico que resulta na formação da imagem apresentando distribuições de cores de acordo com o intervalo de temperatura do corpo. Assim, em tecidos cancerosos há um aumento de temperatura que os diferencia de tecidos sadios, realçando sua localização e indicando a presença da doença. A termografia se diferencia das demais modalidades

pois não analisa as alterações anatômicas. Nesta técnica são observadas as alterações metabólicas e vasculares que se iniciam antes do aparecimento do tumor. Pode ser utilizada tanto para a análise das variações térmicas que ocorrem nas mamas, quanto nas demais partes do corpo e não oferece riscos à saúde.

2.4.1 Mamografia e Equipamento Mamográfico

Mamografia é um tipo específico de exame radiográfico que tem como finalidade registrar imagens da mama, provendo a visualização das estruturas internas e possíveis lesões para auxiliar médicos na descoberta do câncer.

Sua realização é indicada pelo Instituto Nacional do Câncer (INCA) e pela Sociedade Americana de Câncer (ACS) à mulheres entre 40 e 69 anos, e a partir dos 35 anos à mulheres que apresentam câncer de mama no histórico familiar ou que são portadores de predisposição genética, visto que os riscos de aparecimento da doença nesses casos são maiores.

A mamografia continua sendo a técnica mais utilizada e indicada na detecção prévia de lesões não palpáveis na mama (KIM; YOON, 2007); (CALAS et al., 2011). Com a mamografia é possível diagnosticar casos de câncer de mama em fase inicial o que aumentam as chances (até 90%) de ser tratável e reduz a mortalidade por essa doença. Dependendo da qualidade da mamografia é possível rastrear massas a partir de 1 mm, enquanto no auto-exame da mama só é possível perceber massas a partir de 1,5 cm (LEITE et al., 2011).

O exame mamográfico na identificação de tumores apresenta até 88% de sensibilidade e até 99% de especificidade (INCA, 2002). A sensibilidade é a capacidade de detectar verdadeiro positivo, ou seja, em casos em que existe a doença, ela é detectada. Já a especificidade está relacionada à detecção de verdadeiro negativo, em outras palavras significa a detecção correta de casos livres de anomalias. A sensibilidade depende de fatores como tamanho e localização da lesão, densidade do tecido mamário, qualidade dos equipamentos e experiência do especialista (INCA, 2002). Já a especificidade é influencidada pela qualidade da imagem.

Além da sensibilidade e especificidade também existem as taxas que medem o erro no diagnóstico do câncer de mama: falso-positivo (caso em que a lesão é classificada/diagnosticada erroneamente como maligna) e falso-negativo (acontece quando a lesão classificada erroneamente como benigna), o qual pode ocorrer em 10% dos exames.

A leitura de mamografias normalmente é realizada por radiologistas. Essa tarefa demanda formação e muita experiência do profissional, que devido a fatores como má qualidade da imagem, tamanho e variação morfológica das lesões, nem sempre propiciam uma avaliação precisa e uniforme. Estima-se que a sensibilidade dos radiologistas no rastreio do câncer da mama varia entre 65% e 75% (SKAANE; ENGEDAL; SKJENNALD, 1997).

O exame é realizado pelo mamógrafo, aparelho de raio-X que pode ser convencional (analógico) ou digital (Figura 2.5).

Figura 2.5: Mamógrafo: (a) analógico e (b) digital
Fonte: http://www.siemens.com

O mamógrafo convencional utiliza o filme detector que capta os raios-X que atravessam o tecido mamário, exibe e armazena a imagem gerada. O filme é revelado, podendo ser digitalizado e visualizado/analisado pelo radiologista ou médico especialista pela tela do computador. Esta abordagem possui algumas limitações, dentre elas, a falta de padronização na qualidade da imagem, a demora no tempo de processamento da imagem e imutabilidade da mesma. Caso a imagem não apresente o contraste desejado é necessário a repetição do procedimento e uma nova exposição do paciente.

Por outro lado, o mamógrafo digital possui um detector digital que converte a radiação X recebida em sinal elétrico, em seguida esse sinal é quantizado e convertido em sinal digital pelo conversor analógico-digital formando assim a imagem. Após a aquisição, a imagem é transferida eletronicamente, são utilizados computadores e software para exibir, ampliar, clarear ou escurecer a imagem digital de acordo com a necessidade. Esta técnica envolve vantagens como rapidez na realização do exame, o paciente sofre menos exposição à radiação, além de

ser possível realizar ajustes para melhorar a imagem.

Os mamográfos em geral dispõem de um compartimento de compressão de acrílico, no qual é posicionada a mama com o objetivo de imobilizar a paciente durante o exame e comprimir a mama para diminuir e uniformizar sua espessura (Figura 2.6). Essa compressão é um requisito fundamental para garantir a qualidade da imagem e, consequentemente, melhorar a visualização das lesões e estruturas internas da mama, uma vez que a compressão adequada diminui o risco de uma imagem tremida e com ruído, além disso, minimiza a radiação dispersa e a exposição do tecido mamário à radiação.

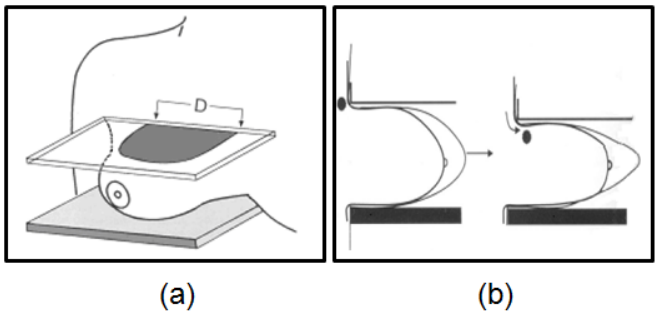

Figura 2.6: Mamográfo: (a) Compartimento de compressão (b) Compressão da mama
Fonte: (KOPANS, 2000)

Os mamógrafos utilizam as seguintes técnicas de posicionamento para explorar a mama (Figura 2.7): Crânio-Caudal (CC); Médio Lateral Oblíqua (MLO); Látero Medial (LM); Médio Lateral (ML); Súpero Inferior Oblíqua (SOI); Lateral Médio Oblíqua (LMO); Crânio-Caudal de Baixo (FB); Ínfero Superior Oblíqua (ISO).

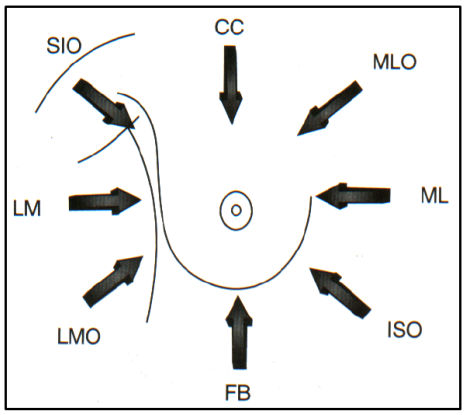

Figura 2.7: Projeções dos feixes de raios X em exames mamográficos
Fonte: (KOPANS, 2000)

Normalmente são realizadas duas incidências em cada mama, a MLO e CC (Figura 2.8), as demais são solicitadas de acordo com o biotipo do(a) paciente e a necessidade da visualização de detalhes das lesões. A MLO é a incidência mais eficaz pois é possível visualizar toda a mama em uma única imagem, enquanto a CC completa a MLO provendo a visualização do parênquima mamário, exceto dos tecidos da região axilar.

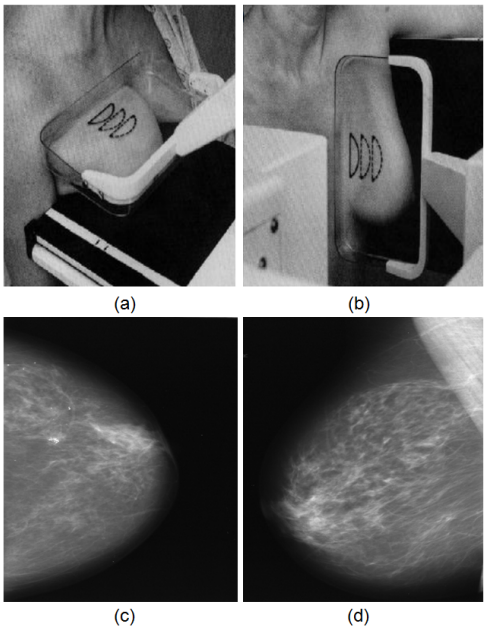

Figura 2.8: Incidência: (a) CC (b) MLO Fonte: (KOPANS, 2000); Mamografia: (c) CC (d) MLO

A formação da imagem de mamografia, em ambos equipamentos (analógico e digital), depende dos diferentes graus de densidade dos tecidos mamários e, consequentemente, da capacidade do tecido de absorver a radiação. Assim, quanto menor a densidade mamária, menor será a exposição de raios-X, pois penetram facilmente em tecidos adiposos fazendo com que estes apareçam escuros (radiolúcido). Já nos tecidos conjuntivo e glandular que são radiologicamente densos (radiodenso), os raios-X não penetram bem, resultando em uma imagem de baixo contraste com áreas predominantemente brancas (radiodenso) (BOYD et al., 1998); (ELSHINAWY; ABDELMAGEED; CHOUIKHA, 2011). Na Figura 2.9, tem-se um exemplo de mamografia digital onde é possível visualizar os principais tecidos que compõe a mama.

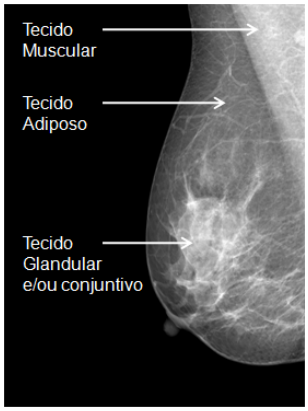

Figura 2.9: Visualização dos diferentes tipos de tecido na mamografia
Fonte: (BCDR, 2012)

2.4.2 Lesões Encontradas em Mamografias

Por meio da mamografia pode-se visualizar as diversas anomalias (assimetria entre as mamas, distorção arquitetural, aumento da densidade do tecido mamário, massas (tumores) e calcificações da mama. Dentre elas, cerca de 80-85% são uma massa, um conjunto de calcificações, ou uma combinação de ambos (Figura 2.10) (MCKENNA, 1994); (KOPANS, 2000).

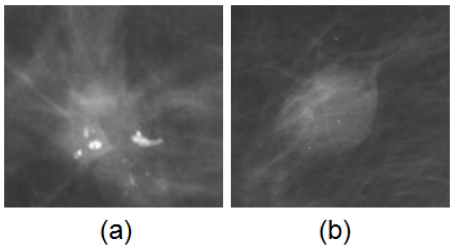

Figura 2.10: Lesões: (a) microcalcificações agrupadas (b) massa
Fonte: (BCDR, 2012)

Grande parte das informações utilizadas pelos radiologistas no diagnóstico é obtida a partir do aspecto visual dos achados mamográficos, que podem ser diferenciados entre si pelo seu tamanho, densidade, número, forma e margem, por exemplo. Em determinados casos também é possível direfenciá-los quanto ao grau de suspeição de malignidade.

Calcificações são depósitos de cálcio na mama que aparecem com o avançar da idade. Em sua maioria, as calcificações são benignas, porém, não se pode descartar a probabilidade de corresponder à processos malignos (LIBERMAN et al., 1998). As mamas podem apresentar simultaneamente calcificações benignas e malignas. Características como forma, distribuição,

número e tamanho auxiliam nesta distinção.

As calcificações podem ser cutâneas, vasculares, com formato redondo, linear, ramificado, espiculado, etc. Quando apresentam formato indefinido (amorfo), espiculado, linear ou ramificado possuem maior probabilidade de malignidade.

Quando menores que 0,5 mm são definidas como microcalcificações, já as partículas maiores que 0,5 mm recebem o nome de macrocalcificações (PAULINELLI; CALAS; FREITAS JUNIOR, 2007). Quanto menor a partícula de calcificação, maiores são as chances de ser maligna. Partículas com diâmetro superior a 3 mm tendem a ser benignas.

Elas também podem apresentar-se distribuídas de maneira difusa, regional, agrupadas, lineares e segmentares (ACR, 2013b). Calcificações (mais de 3) agrupadas são o primeiro e importante sinal de câncer de mama (GARUD; SHAHARE, 2013). Quando detectadas agrupadas e sem associação à massas, a probabilidade de ser maligna é cerca de 20-35%. Assim como, dispostas de forma linear e segmentares também sugerem malignidade. Já as calcificações com distribuíção regional e difusa geralmente são benignas.

Em relação à massa, pode-se definir como uma estrutura tridimensional que ocupa espaço e pode ser vista em pelo menos duas projeções mamográficas diferentes (MLO/CC) (ACR, 2003). Quando visualizada apenas em uma projeção é chamada de assimetria.

Dessa forma, os radiologistas devem considerar os seguintes parâmetros para a classificação das massas (ACR, 2013b):

- Tamanho - dependo do tamanho da massa, ela pode ser detectada por meio de exames clínicos (palpável), em casos de massas muito pequenas elas só podem ser identificadas através da mamografia (não palpável). Uma vez que massa é palpável, a mamografia apresenta taxa de falso negativo de 10-15%. O ideal é que a massa seja identificada quando menor que 2 mm, porém, é difícil identificar tumores menores que 5 mm;

- Forma - as massas podem apresentar formato redondo, oval ou irregular (Figura 2.11). A morfologia irregular da massa muitas vezes indica a não uniformidade do seu crescimento. Assim, ao visualizar a Figura 2.11, nota-se que a malignidade está relacionada ao formato irregular do tumor.

- Margem - é o termo dado à relação da massa com o tecido circunvizinho. É um dos fatores

mais importantes na determinação da benignidade/malignidade de uma lesão (KOPANS, 2000). Contornos espiculados ou mal definidos indicam que uma lesão está invadindo o tecido adjacente, o que sugere malignidade, como é possível observar na Figura 2.12. Por outro lado, lesões como fibroadenoma ou cisto que apresentam forma e margem bem definidas, são benignas.

De acordo com a *American College of Radiology (ACR)* e o Colégio Brasileiro de Radiologia (CBR), existem cinco tipos de margens:

- Circunscrita - massa benigna geralmente possui margem nítida que a delimita do tecido circunvizinho;
- Obscurecida - a margem é escondida pelo tecido circunvizinho normal;
- Microlobulada - a margem apresenta ondulações na superfície que podem caracterizar o câncer de mama;
- Indistinta - massas malignas possuem margem mal definidas devido à invasão aos tecidos circunvizinhos;
- Espiculada - projeções fibrosas se estendem da massa principal caracterizando-a como maligna.

- Densidade - também é possível classificar a densidade da massa em relação ao tecido glandular normal circunvizinho. Essa classificação se dá em quatro grupos: alta densidade, isodenso (densidade igual), baixa densidade e contendo gordura. A maioria das massas mamárias são isodensa ou de alta densidade. Sendo assim, as massas que apresentam alta densidade (maior que a densidade do tecido glandular) tendem a ser malignas (TAHMASBI; FATEMEHSAKI; SHOKOUHI, 2011).

O terceiro achado mamográfico mais comum em mamografias e sugestivo de malignidade é a distorção arquitetural. Essa anomalia pode ser definida como a distorção da arquitetura normal do tecido mamário, não apresenta massa visível e inclui espiculações irradiando a partir de um ponto central (Figura 2.13) ou a distorção na borda do parênquima (ACR, 2003).

As distorções podem indicar a presença do câncer de mama quando associadas à uma massa, assimetria ou calcificações. São benignas quando se trata de um erro de posicionamento da

mama durante a realização da mamografia que resulta na sobreposição de tecido normal, ou de uma cicatriz pós cirurgia/biopsia. Para a confirmação do diagnóstico, elas devem ser analisadas cuidadosamente por meio de outras incidências, além da MLO e CC, e em determinados casos deve-se realizar a biópsia.

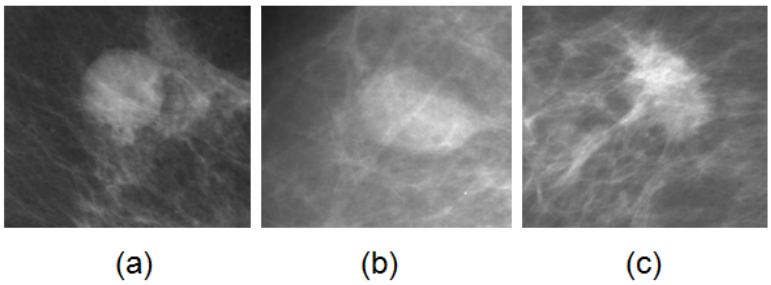

Figura 2.11: Morfologia dos tumores: (a) redondo (b) oval (c) irregular
Fonte: (BCDR, 2012)

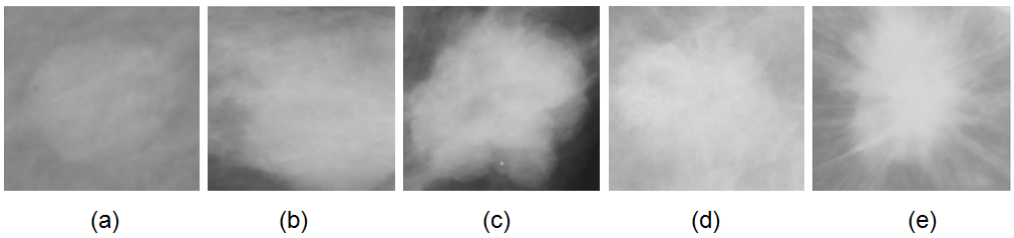

Figura 2.12: Classificação das margens: (a) circunscrita (b) obscurecida (c) microlobulada (d) indistinta (e) espiculada
Fonte: (BCDR, 2012)

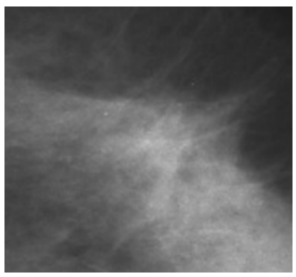

Figura 2.13: Distorção arquitetural
Fonte: (BCDR, 2012)

2.4.3 BI-RADS

No ano de 1993, o sistema *Breast Image Reporting and Data System (BI-RADS)* foi aprovado pelo *ACR* com a colaboração dos comitês *American Medical Association*, *American College*

of Surgeons, *Centers for Disease Control and Prevention*, *College of American Pathologistis*, *College of Surgeons*, *Food and Drug Administration* e *National Cancer Institute*, visando a padronização da nomenclatura de laudos mamográficos para auxiliar médicos no diagnóstico quanto à probabilidade da lesão ser maligna e propondo uma conduta específica para cada caso (ACR, 2003).

Esse sistema é amplamente utilizado no mundo todo como uma linguagem padrão entre mastologistas, radiologistas e ginecologistas para evitar ambiguidades ou distorções de interpretação de laudos mamográficos.

O BI-RADS foi concebido inicialmente como uma ferramenta mamográfica, mas em 2003 sofreu uma atualização e foi disponibilizado também para a ultrassonografia e ressonância magnética. Sua versão mais atual (5° edição), foi lançada em 2013 e encontra-se estruturada nas seguintes seções:

1. Léxico da imagem mamográfica - os achados mamográficos são descritos de acordo com o léxico, que trata-se da atribuição de um termo específico que descreve características de massas (forma, margens e densidade), calcificações (morfologia e distribuição), distorção de arquitetura, entre outras anomalias. Grande parte da terminologia desenvolvida para a mamografia também é utilizada para os dois novos léxicos (ultrassonografia e ressonância magnética), além de termos próprios referentes às características de cada modalidade, não abordados neste trabalho. Na Tabela 2.1 estão descritas as características das principais lesões (massas e calcificações) encontradas em mamografias;

2. Sistematização do laudo mamográfico - sistema de comunicação projetado para fornecer uma abordagem organizada para a interpretação da imagem e organização dos relatórios. O laudo deve ser organizado de maneira que seja informado se o exame sob análise foi comparado à mamografias anteriores, também deve especificar a composição (densidade) das mamas (Tabela 2.2), o tipo de lesão encontrada (massa, calcificações) e a classificação do exame dentre as categorias estabelecidas (negativo, benigno, provavelmente benigno, suspeito, altamente sugestivo de malignidade), de acordo com a suspeição dos achados;

3. Acompanhamento e monitoramento do resultado final - seção sobre auditoria de mamografia que descreve os dados a serem coletados e utilizados para calcular dados importan-

tes permitindo a cada radiologista avaliar seu desempenho na interpretação de mamografias;

4. Criação de um banco de dados nacional - o resultado padronizado da classificação de imagens de acordo com o BI-RADS é colocado em um banco de dados para ajudar radiologistas no diagnóstico cada vez mais preciso;

5. Orientação do capítulo - ao longo dos anos de uso continuado do BI- RADS, o comitê tem recebido perguntas e relatos de problemas relacionados às várias seções do BI-RADS. Portato, nessa seção são explicadas as razões para determinadas inclusões e mudanças ocorridas no léxico.

Tabela 2.1: Terminologia BI-RADS. Fonte: (ACR, 2013b)

Avaliação Mamográfica	Características	
Massa	Morfologia	Oval
		Redonda
		Irregular
	Margem	Circunscrita
		Obscurecida
		Microlobulada
		Indistinta
		Espiculada
	Densidade	Alta densidada
		Isodenso
		Baixa densidade
		Contém gordura
Calcificações	Tipicamente Benigna	Cutânea
		Vasculares
		Grosseiras (semelhantes a “pipoca”)
		Bastonetes longos
		Redondas

		Em “Casca de ovo” ou em “anel”
		Distróficas
		Leite de cálcio
		Fios de sutura
	Suspeita	Amorfa
		Heterogêneas grosseiras
		Pleomórficas
		Lineares ou Ramificações lineares
		Espiculada
	Distribuição	Difusa
		Regional
		Agrupadas
		Lineares
		Segmentares

Tabela 2.2: Terminologia BI-RADS composição mamária. Fonte: (ACR, 2013b)

Composição Mamária	
BI-RADS 1	Mamas são quase inteiramente adiposas
BI-RADS 2	Existem áreas espalhadas de densidade fibroglandular
BI-RADS 3	As mamas são heterogeneamente densas, o que pode obscurecer pequenas massas
BI-RADS 4	As mamas são extremamente densas, o que reduz a sensibilidade da mamografia

Para a finalização do laudo, é avaliado o risco de malignidade da lesão e realizada a classificação do padrão mamográfico em uma das seguintes categorias BI-RADS (ACR, 2013a):

- BI-RADS 0 (Inconclusivo) - representa um estudo incompleto, onde é necessário solicitar outro exame de imagem (ultrassonografia, ampliação, visão especial de mamografia) para complementar e/ou comparar com exames prévios;
- BI-RADS 1 (Normal) - a mamografia é considerada negativa, as mamas são simétricas e

não apresentam massa ou calcificação. O exame deve ser repetido anualmente ou bianualmente;

- BI-RADS 2 (Benigno) - a mamografia é negativa para neoplasias malignas, ou seja, não há indícios de câncer, porém, estão incluídos nesta categoria achados com 0,05% de malignidade como linfonodos intramamários, fibroadenomas calcificados, cistos mamários simples, calcificações múltiplas de origem secretória, cistos oleosos, lipomas, galactoceles e/ou hamartomas de densidade mista. O acompanhamento deve ser anual ou bianual;

- BI-RADS 3 (Provavelmente benigno) - apresenta lesões com características radiográficas benignas, com menos de 2% de chances de malignidade, porém, é recomendado o acompanhamento em intervalos periódicos de 4 a 6 meses para avaliar a evolução das lesões. Nesta categoria podem ser encontradas massa sólida circunscrita não calcificada, assimetrias focais que diminuem ou desaparecem à compressão e/ou grupamentos de calcificações puntiformes;

- BI-RADS 4 (suspeito) - os achados mamográficos não apresentam aparência clássica de malignidade, porém, apresentam real probabilidade de serem malignas. Para a conclusão do diagnóstico é indicada a investigação citológica ou histológica. Esta categoria encontra-se dividida em 3 subcategorias:

  - 4A (Baixa suspeita) - lesões com 2-10% de chance de malignidade. A mama pode apresentar cistos que necessitam de aspiração, massas palpáveis sólidas e/ou massas com margens parcialmente circunscritas;
  - 4B (Intermediária suspeita) - lesões com 10-50% de chance de malignidade. Pode-se encontrar massas de margens indistintas e com algumas áreas circunscritas;
  - 4C (Moderada suspeita) - lesões com 50-95% de chance de malignidade. Pode apresentar massas irregulares, mal-definidas ou grupamentos de calcificações pleomórficas;

- BI-RADS 5 (Altamente suspeito) - a mamografia apresenta lesões com características acima de 95% de malignidade, como massas espiculadas, de alta densidade, calcifica-

ções lineares finas e/ou massas espiculadas com calcificações pleomórficas associadas. É necessário a realização da biópsia para a confirmação.

- BI-RADS 6 (Maligno) - lesão com 100% de malignidade comprovada por meio do resultado da biópsia. Lesão não retirada ou tratada.

2.5 DIAGNÓSTICO AUXILIADO POR COMPUTADOR

Estudos apontam que limitações próprias do radiologista como fadiga, desatenção e falta de experiência influenciam para que cerca de 10% a 30% das lesões mamárias não sejam identificadas em exames de rotina (CALAS; GUTFILEN; PEREIRA, 2012). Além disso, fatores como a estrutura complexa radiográfica da mama, a similaridade existente entre a massa e tecido glandular, as variações morfológicas das lesões e a baixa qualidade da imagem têm contribuído para interpretações falso-negativas da mamografia pelo radiologista.

Uma maneira de amenizar esse problema e aumentar a sensibilidade na detecção do câncer de mama é a dupla leitura da mamografia, em que dois ou mais especialistas avaliam independentemente as imagens do mesmo exame e realizam o diagnóstico. É comprovado que a dupla leitura aumenta em até 8,5% a detecção do câncer de mama (BENVENISTE; FERREIRA; AGUILLAR, 2006).

Em casos que não se pode realizar a dupla leitura, ou que existe alta variabilidade de opinião interobservadores, ou até mesmo a falta de profissionais devidamente treinados, é indicada a utilização de sistemas de diagnóstico auxiliado por computador (CAD) (CALAS; GUTFILEN; PEREIRA, 2012). Esses sistemas funcionam como um leitor de exames por imagens para auxiliar radiologistas na interpretação e na tomada de decisão, fornecendo-lhes uma segunda opinião sobre o diagnóstico.

De acordo com Doi (2007), as primeiras técnicas computacionais para auxiliar na análise de imagens médicas foram elaboradas a partir de 1960, mas só na década de 1980 foi introduzido o conceito de sistemas de diagnóstico auxiliado por computador. Entretanto, só nos últimos anos tem-se observado a intensificação no desenvolvimento de sistemas CAD nas diversas áreas da medicina.

No diagnóstico auxiliado por computador, o radiologista utiliza informações resultantes da análise quantitativa automatizada de imagens (de radiografia, tomografia computadorizada,

ressonância magnética ou ultrassonografia) que podem ser de diversas partes do corpo, como crânio, tórax, abdome, osso ou sistema vascular para realizar o diagnóstico (MARQUES, 2001).

Os CADs para mamografia são desenvolvidos especialmente para a detecção de massas e/ou microcalcificações, visto que são as lesões mais incidentes na mama. Eles podem fornecer como resposta a probabilidade da lesão ser maligna ou benigna. Sua sensibilidade para a detecção de calcificações varia entre 80% a 100%, já para a detecção de massas a sensibilidade é menor, de 88% até 92% (MORTON et al., 2006) (BAUM et al., 2002). A dificuldade na detecção de massas se dá principalmente devido à sua similaridade com o tecido glandular, pois o mesmo obscurece a sua visualização.

Dessa forma, o objetivo no desenvolvimento dos CADs tem sido o aumento da sensibilidade no rastreio mamográfico. Porém, como consequência tem-se também o aumento do número de falsos-positivos, o que pode degradar significativamente a eficácia do CAD (LI et al., 2006). Isso significa que o diagnóstico final cabe sempre ao especialista e que é importante a reavaliação das respostas do CAD, mesmo que por isso o processo de tomada de decisão se torne mais lento.

Então, nota-se que a sensibilidade do CAD pode variar de acordo com: a composição mamária; a interpretação do radiologista; e as funcionalidades (métodos) que o sistema dispõe. Em geral, os CADs otimizam a qualidade da imagem, facilitando a visualização/interpretação das lesões e/ou fornecem uma resposta (diagnóstico) que pode ser utilizada como referência, contribuindo para aumentar a eficácia do diagnóstico.

Quanto às funcionalidades e especificidades dos sistemas CAD, existem dois tipos de aplicações:

- Detecção auxiliada por computador (do inglês *Computer-Aided Detection (CADe)*) - sistema que auxilia na identificação de lesões (regiões de interesse) a partir de padrões radiológicos suspeitos por meio da varredura da imagem pelo computador. Tem a finalidade de fornecer avisos visuais na mamografia indicando regiões suspeitas que merecem atenção e investigação (BOZEK; DELAC; GRGIC, 2008). É útil em detectar regiões contendo lesões que muitas vezes poderiam ser negligenciadas pelo radiologista em uma única leitura. CADe não inclui a classificação automatizada da lesão, que neste caso é realizada apenas pelo radiologista;

- Diagnóstico auxiliado por computador (do inglês *Computer-Aided Diagnosis (CADx)*) - sistema responsável por classificar a lesão investigada no padrão normal ou anormal, a partir de informações extraídas da imagem. CADx não substitui o radiologista, pelo contrário, trata-se de uma ferramenta de apoio à decisão que pode aumentar a sensibilidade do radiologista no rastreio do câncer de mama em até 21% (CALAS; GUTFILEN; PEREIRA, 2012). O radiologista combina dados do histórico clínico do paciente com informações fornecidas pelo CAD para o diagnóstico final.

Em geral, os sistemas CAD abrangem técnicas provenientes de duas áreas do conhecimento (MARQUES, 2001) que serão abordadas no próximo capítulo:

- Visão computacional - utiliza técnicas de processamento digital de imagem para realçar as lesões na imagem, segmentar (isolar) a região de interesse e extrair atributos;
- Inteligência artificial - inclui métodos de seleção de atributos mais representativos, métodos estatísticos e reconhecimento de padrões para a classificação das lesões.

2.6 MEDIDAS DE DESEMPENHO

Existem vários índices de desempenho que podem ser utilizados para avaliar os resultados da classificação realizada por sistemas de auxílio ao diagnóstico. A matriz de confusão (*confusion matrix*), por exemplo, apresenta os resultados de uma classificação sob forma de uma tabela de duas classes: classe verdadeira versus classe predita pelo modelo. Os números localizados na diagonal principal da matriz representa a quantidade de acerto e os demais elementos representam os erros da classificação (Tabela 2.3). É importante esclarecer que a matriz de confusão é apresentada neste trabalho baseando-se em exemplos de problemas com apenas duas classes (benigno e maligno), porém, pode ser aplicada a várias classes.

Tabela 2.3: Uma matriz de confusão para um problema com duas classes.

	Posivos previstos	Negativos previstos
Positivos	VP	FP
Negativos	FN	VN

Onde:

- Verdadeiro Positivo (VP) corresponde ao número de verdadeiros positivos, ou seja, o número de exemplos da classe positiva classificados corretamente;
- Falso Positivo (FP) corresponde ao número de falsos positivos, ou seja, o número de exemplos cuja classe é negativa mas que foram classificados incorretamente como pertencente à classe positiva;
- Verdadeiro Negativo (VN) corresponde ao número de verdadeiros negativos, ou seja, o número de exemplos da classe negativa classificados corretamente;
- Falso Negativo (FN) corresponde ao número de falsos negativos, ou seja, o número de exemplos pertencentes à classe positiva mas foram preditos incorretamente como da classe negativa.

A partir da matriz de confusão pode-se derivar outras medidas, como é possível observar na Tabela 2.4.

Tabela 2.4: Medidas de desempenho.

Medida	Equação	Descrição
Sensibilidade (S)	$\frac{VP}{(VP+FN)}$	Taxa de detecção de verdadeiros positivos.
Especificidade (E)	$\frac{VN}{(VN+FP)}$	Taxa de detecção de verdadeiros negativos.
Precisão ou Valor Preditivo Positivo (VPP)	$\frac{VP}{(VP+FP)}$	Proporção de exemplos positivos classificados corretamente entre os preditos como positivos.
Valor Preditivo Negativo (VPN)	$\frac{VN}{(VN+FN)}$	Proporção de exemplos negativos classificados corretamente entre os preditos como negativos.
Taxa de Falso Positivo (α)	$1-E$	Exemplos da classe negativa classificados incorretamente como positivo.

Taxa Falso Negativo (β)	$1 - S$	Exemplos da classe positiva classificados incorretamente como negativo.
F-measure	$2\left(\frac{Precisao*S}{Precisao+S}\right)$	Mede a eficácia de um classificador. Trata-se da média harmônica entre precisão e recall.
Acurácia (AC)	$\frac{(VP+VN)}{(VP+VN+FP+FN)}$	Corresponde à taxa de exemplos positivos e negativos classificados correctamente.

2.7 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foram descritos conceitos fundamentais para a compreensão dos termos relacionados ao câncer de mama e imageamento mamográfico. Também foi apresentada a diferença entre os sistemas de diagnóstico auxiliado por computador e sistemas de detecção por computador, bem como medidas de desempenho para avaliar o desempenho de tais sistemas. No próximo capítulo serão apresentados os demais conceitos para o embasamento teórico deste trabalho, assim como técnicas de processamento digital de imagens e mineração de imagens.

CAPÍTULO 3

MINERAÇÃO E PROCESSAMENTO DE IMAGENS

Neste capítulo são abordadas as etapas constituintes do processo de descoberta de conhecimento em banco de dados, da mineração de imagens e do processamento digital de imagens. Além disso, é realizada uma revisão bibliográfica de trabalhos que apresentaram como objetivo o desenvolvimento de uma metodologia CAD para o diagnóstico do câncer de mama.

3.1 PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS - DCBD

Descoberta de Conhecimento em Banco de Dados - DCBD (do inglês *Knowledge Discovery in Databases (KDD)*) pode ser definido como "um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados" (FAYYAD; SHAPIRO; SMYTH, 1996).

Neste trabalho consideramos a divisão do processo de DCBD em três grandes etapas principais: pré-processamento, mineração de dados (MD) e interpretação/avaliação de padrões (Figura 3.1). As mesmas serão detalhadas respectivamente nas subseções: 3.1.1;3.1.2 e 3.1.3.

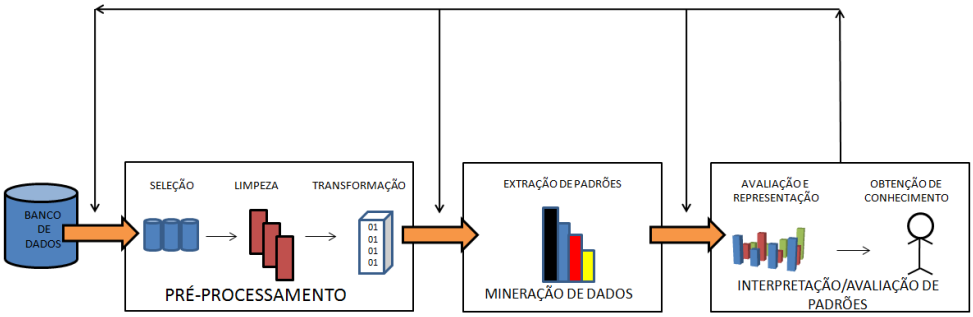

Figura 3.1: Etapas DCBD

Resumidamente, pode-se definir o pré-processamento como uma etapa composta por tarefas que visam garantir a qualidade dos dados que passarão pelo processo de DCBD. Na mineração de dados são aplicadas técnicas para extrair padrões e relacionamentos existentes entre os dados. As informações geradas na MD são analisadas, avaliadas e interpretadas na última etapa do

DCBD, interpretação/avaliação de padrões.

Por se tratar de um processo iterativo, se o conhecimento obtido não for de interesse, pode-se retornar às etapas anteriores e repetir o processo realizando-se alterações nos parâmetros, incluindo ou eliminando dados, por exemplo.

O DCBD tem interdisciplinaridade com diversas áreas do conhecimento como inteligência artificial (especificamente com a Aprendizagem de Máquina (AM)), reconhecimento de padrões, banco de dados e estatística (FAYYAD; SHAPIRO; SMYTH, 1996). Algoritmos de AM são aplicados sobre conjuntos de dados para a construção de modelos por meio da estatística e da distribuição de probabilidade. Além disso, os padrões descobertos são validados estatisticamente.

3.1.1 Pré-processamento

É comum encontrar nos repositórios dados ruidosos (com erros), incompletos (com ausência de valores), duplicados (mais de um objeto com os mesmos valores para todos os atributos ou mais de um atributo com os mesmos valores para mais de um objeto) e inconsistentes (contradizem valores de outros atributos do mesmo objeto) (FACELI et al., 2011). Esses tipos de dados são originados por meio de uma falha (humana, hardware, software) mediante um processo de coleta, transmissão, armazenamento ou entrada. Como se tratam de dados de baixa qualidade, devem ser eliminados ou corrigidos na etapa de pré-processamento, pois influenciam diretamente na relevância do conhecimento que se deseja descobrir. Devido à sua importância, essa etapa pode consumir até 80% do tempo dedicado ao processo de descoberta de conhecimento (MCCUE, 2007).

As tarefas realizadas no pré-processamento para a preparação dos dados são:

- Seleção - a partir do banco de dados é selecionado o conjunto de dados que passará pelo processo de DCBD;
- Limpeza - garante a qualidade dos dados em relação à completude, veracidade e integridade. Procedimento em que os dados incompletos e ruidosos são corrigidos ou eliminados para que não comprometam o conhecimento obtido ao término do processo;
- Transformação - são aplicadas técnicas para transformar os dados em formatos utilizáveis

para os algoritmos de mineração de dados. Estas são:

- Normalização - em casos que os limites de valores dos atributos são muito discrepantes, podem ser escalonados para uma faixa específica com limites mínimo e máximo. Ex: valor dos atributos na escala de -1.0 até 1.0;
- Integração - quando se tem a necessidade de combinar diferentes bases de dados em uma só;
- Conversão de dados - algumas técnicas de mineração de dados são limitadas à manipulação de determinados tipos (numérico, nominal) de atributos. Essa limitação faz com que seja preciso a realização da transformação de um tipo de dados para outro. Ex: conversão de nominal para inteiro ou de ordinal para binário;
- Agregação - visa a redução do número de atributos (redução da dimensionalidade) para facilitar a manipulação de dados. Pode ser realizada de duas maneiras (FACELI et al., 2011): (1) os atributos originais são combinados e substituídos por novos atributos; (2) uma amostra composta pelos atributos mais importantes é selecionada e os demais são descartados.

3.1.2 Mineração de Dados

A Mineração de Dados (MD) surgiu no final da década de oitenta com o objetivo de extrair informações úteis a partir de um grande conjunto de dados. Mediante esse objetivo, é considerada por diversos autores como similar ao DCBD, porém, pode-se concluir que a MD é apenas uma etapa do processo completo de descoberta de conhecimento em banco de dados (FAYYAD; SHAPIRO; SMYTH, 1996).

A MD dispõe de várias tarefas e sua execução está relacionada ao que se deseja encontrar nos dados, podendo ser regularidades, associações, similaridades, etc. Segundo (FACELI et al., 2011), as tarefas de MD encontram-se divididas em preditivas e descritivas. As tarefas preditivas têm como objetivo prever o valor (rótulo) de um atributo com base em um modelo de dados de treinamento. Assim, os algoritmos preditivos seguem o paradigma de aprendizado supervisionado, pois no treinamento é fornecido uma classe à qual cada amostra pertence (SILVA, 2008). As tarefas preditivas são:

- Classificação - o algoritmo recebe como entrada um conjunto de dados discretos (do tipo nominal) que são mapeados em uma ou várias classes categóricas predefinidas, de modo que esses dados se relacionem à classe;

- Regressão - similar à tarefa de classificação, mas com uma aplicação numérica. Sua função é mapear dados (contínuos e ordenados) com base em um modelo numérico. Diferente da tarefa de classificação, em que os dados pertencem à categorias, na regressão o dado (número) pertence a uma escala.

Por outro lado, a aplicação das tarefas descritivas possibilita a extração de padrões e relacionamentos entre os dados, revela tendências, anomalias, entre outras correlações. Nesta abordagem, a priori não se tem informação a respeito dos dados, o que caracteriza a aprendizagem como não supervisionada. Dentre as tarefas descritivas pode-se encontrar:

- Descoberta de associação - algoritmos geram regras para descobrir correlações (associações) entre um conjunto de dados armazenados. As regras de associação apresentam o seguinte formato: o lado esquerdo é chamado de antecedente (Se) e o lado direito de consequente (Então).

  Exemplo de regra: sexo(X,[fem]) $\Rightarrow$ consome (X,[sapato, bolsa]);

- Agrupamento (*clustering*) - identifica e agrupa objetos que possuem maior similaridade entre si, baseados nos valores dos atributos. O agrupamento visa maximizar a similaridade entre os elementos do grupo (intracluster) e maximizar a diferença entre os grupos (intercluster). O critério de similaridade e o número de grupos é definido pelo analista;

- Sumarização - consiste na utilização de métodos para encontrar e apresentar uma descrição compacta para um subconjunto de dados. Nesta tarefa utiliza-se medidas estatísticas, como média e desvio padrão;

- Detecção de desvios (*outliers*) - diferentemente das outras tarefas de MD, esta identifica padrões com pouca incidência no BD ou que apresentem valores diferentes dos que já foram observados (GOLDSCHIMIDT; PASSOS, 2005);

- Descoberta de sequências - busca por dados que relacionam-se de forma regular em um determinado período de tempo.

3.1.3 Interpretação/Avaliação de Padrões

Na última etapa do processo de descoberta do conhecimento é realizada a interpretação e a avaliação das informações obtidas na etapa de mineração de dados. Essas tarefas geralmente são realizadas por um especialista em DCBD e um especialista da área em estudo. Eles avaliam os resultados por meio de critérios como precisão (exatidão), compreensibilidade (simplicidade), aplicabilidade (onde pode ser útil sua aplicação), interessabilidade (quão interessante/inesperado) e complexidade computacional (tempo de processamento). Nessa etapa também são utilizadas medidas estatísticas e métodos que eliminam resultados inexatos (TAN; STEINBACH; KUMAR, 2009) para garantir e comprovar a validade e veracidade da informação.

Uma vez aplicada medidas de qualidade na avaliação dos resultados é possível saber se o conhecimento obtido é realmente útil e de fácil compreensão pelos usuários. Um modelo de fácil compreensão está relacionado, por exemplo, ao número de regras e de condições por regra geradas pelo algoritmo, visto que, a complexidade da interpretação do modelo é proporcional a esses fatores.

3.2 MÉTODOS DE MINERAÇÃO DE DADOS

Para cada tarefa de mineração de dados existe um conjunto de técnicas associadas. A técnica diz respeito a qualquer teoria que fundamenta a implementação de um método (GOLDSCHIMIDT; PASSOS, 2005). Existe uma grande variedade de métodos de mineração, mas neste trabalho vamos focar nos métodos de classificação supervisionada como árvore de decisão, KNN e redes neurais.

3.2.1 Árvore de Decisão (*Decision Tree*)

A Árvore de Decisão (AD) é um método de aprendizagem popularmente utilizado para realização de inferência indutiva. É aplicada com sucesso em tarefas de aprendizagem, como classificação e regressão, inclusive para o diagnóstico de casos médicos (MITCHELL, 1997).

"Uma árvore de decisão toma como entrada um objeto ou situação descritos por um conjunto de atributos e retorna uma decisão - valor de saída previsto, de acordo com a entrada" (RUSSELL; NORVIG, 2003).

Seu funcionamento consiste em aplicar uma estratégia para dividir recursivamente um problema complexo em subproblemas e, em seguida, combinar as soluções encontradas em formato de uma árvore para produzir uma solução final do problema (FACELI et al., 2011).

Considerando que uma árvore de decisão é uma fonte de informação, se faz necessário a utilização de uma medida da Teoria da Informação, chamada entropia, para aumentar o Ganho de Informação (GI) ao término do processo de aprendizagem.

Entende-se que entropia mede a aleatoriedade (dificuldade para predizer) do atributo que define as classes. Assim, o atributo do nó de divisão é escolhido de maneira que a entropia seja minimizada, assim como, o tamanho da árvore para tornar o modelo mais compreensível pelo usuário.

O cálculo da entropia de um conjunto de dados S com i classes se dá pela Equação 3.1:

$$Entropia(S) \equiv \sum_{i=1}^{n} -p_i log_2(p_i) \tag{3.1}$$

Seja:

- n_i o número de amostras da classe i;
- n a quantidade total de amostras;
- $p_i = \frac{n_i}{n}$ corresponde a probabilidade de uma amostra pertencer a classe i.

Na continuidade é calculado o ganho de informação que se dá pela redução na entropia esperada mediante o particionamento dos dados de acordo com o atributo do nó de divisão (MITCHELL, 1997). Ou seja, pode-se medir quão eficaz um atributo é no particionamento de um conjunto de dados em suas respectivas classes. Esse cálculo se dá pela Equação 3.2:

$$Ganho(S,A) \equiv Entropia(S) - \sum_{v \in valores(A)} \left(\frac{S_v}{S}\right) Entropia(S_v) \tag{3.2}$$

O primeiro termo da Equação 3.2 refere-se à entropia do conjunto S. O segundo termo trata-se do valor da entropia de S depois do particionamento das amostras de S por um atributo A. A entropia esperada descrita no segundo termo é referente à soma das entropias de cada subconjunto de S_v, ponderado pela fração de exemplos que pertencem a S_v (MITCHELL, 1997).

Seja:

- valores(A) o conjunto dos possíveis valores para o atributo A; e
- S_v o subconjunto de S para o qual o atributo A tem valor v.

Na Figura 3.2 é mostrado um exemplo de árvore de decisão que tem como objetivo determinar quais dias são favoráveis para jogar tênis. Os atributos que maximizaram o ganho de informação e, consequentemente, utilizados nos nós de divisão foram "Clima", "Vento" e "Umidade".

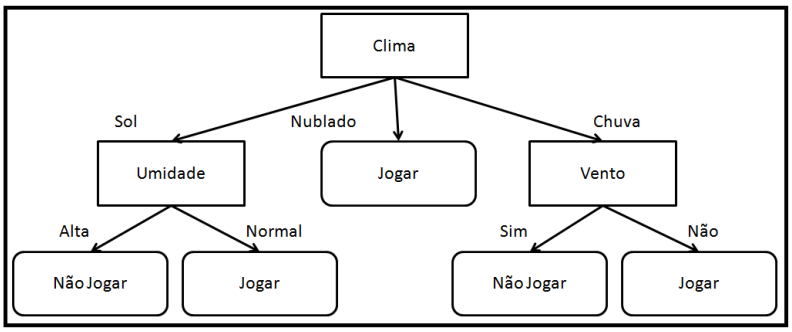

Figura 3.2: Árvore de decisão
(MITCHELL, 1997)

As árvores de decisão podem ser representadas por um conjunto de regras disjuntas. Onde cada caminho da árvore, iniciando na raiz até o nó folha, corresponde a uma regra de decisão. No caso da AD da Figura 3.2, o conjunto de regras correspondente ao dia satisfatório para jogar tênis é:

$(Clima = Sol \wedge Umidade = normal) \vee (Clima = nublado) \vee (Clima = Chuva \wedge Vento = Nao)$

Existem vários algoritmos de árvore de decisão. Os mais bem-sucedidos e representativos de são: ID3 (QUINLAN, 1986), C4.5 (QUINLAN, 1993) e CART (BREIMAN et al., 1984).

O algoritmo *Iterative Dichotomiser (ID3)*, proposto por Quinlan, foi um dos primeiros algoritmos de aprendizagem por árvore de decisão. O ID3 constrói a AD a partir de conjunto de treinamento e a utiliza como modelo para a classificação de novas amostras. Para a realização da inferência, o ID3 se baseia nos fundamentos de ganho de informação e entropia.

O ID3 possui uma versão aprimorada, chamada de C4.5. Essa versão apresenta algumas modificações como a manipulação de atributos contínuos e com valores faltantes, também possibilita determinar a profundidade da árvore aplicando o método de poda (MITCHELL, 1997).

A poda consiste em remover uma subárvore enraizada em um determinado nó, tornando-o nó folha (MITCHELL, 1997). Esse procedimento diminui o tamanho da árvore e pode ocorrer durante sua construção (pré-poda) a partir de um critério de parada ou após a árvore ter sido construída (pós-poda).

O C4.5 também apresenta uma versão em Java chamada de J48, e um sucessor, o See5/C5. Comparado ao C4.5 o algoritmo C5 é mais rápido computacionalmente, utiliza menos memória, apresenta resultados mais precisos, gera árvores de decisão menores, entre outras melhorias.

Outro algoritmo baseado em árvore muito utilizado é o *Classification and Regression Trees (CART)*. Desenvolvido por (BREIMAN et al., 1984), o CART pode ser aplicado tanto à problemas de classificação quanto de regressão. Quando aplicado à regressão é chamada de árvore de regressão.

Em geral, os algoritmos de indução de árvore de regressão são semelhantes aos algoritmos de árvore de decisão, inclusive, a interpretação dos modelos gerados também é igual. O que os difere é que na árvore de regressão, cada nó folha contém um valor numérico referente a uma média (ou equação) de todos os valores do conjunto de treinamento, ou seja, um valor preditivo, enquanto na classificação temos para cada nó folha um modelo linear (chamado de classe) que retorna uma decisão. É importante lembrar que o método de poda também se aplica à árvores de regressão, neste caso os ramos candidatos a poda são os que apresentam menor poder de predição.

O processo de divisão da árvore procura minimizar o erro quadrático (a variância da amostra multiplicada pelo número de casos) em cada nó. Assim, os atributos de divisão são escolhidos baseados no índice Gini que tem por base o cálculo da entropia (Equação 3.1). Esse índice mede o grau de impureza de um conjunto de dados (diz-se que o conjunto é puro quando todos seus elementos pertencem à mesma classe). Portanto, o atributo que resultar na maior diminuição da impureza é selecionado para o nó de divisão. O índice Gini é obtido pela Equação 3.3:

$$G(t) = 1 - \sum_{i=1}^{c} p_i^2 \tag{3.3}$$

Onde:

- t é o identificador de cada nó da árvore;

- $G(t)$ é o valor esperado da soma dos erros quadráticos da regressão;
- p_i é a frequência relativa de cada classe em cada nó; e
- c é o número de classes.

3.2.2 Vizinho mais Próximo (*Nearest Neighbour*)

O vizinho mais próximo trata-se de um método que classifica uma instância com base na(s) instância(s) mais próxima(s)/similar(es) a ela (FACELI et al., 2011). Sua utilização é apropriada principalmente quando os valores dos atributos são contínuos. O algoritmo baseado neste método é o *K-Nearest Neighbour (K-NN)*, onde K representa o número de vizinhos (instâncias) mais próximos (similares) considerados na classificação.

O conjunto de treinamento da Tabela 3.1 é utilizado como modelo para classificar a nova instância da Tabela 3.2. Na Figura 3.3 é possível observar a distribuição dos dados no espaço bidimensional. As instâncias positivas são representadas pelo simbolo "+" e as instâncias negativas pelo "-". O círculo representa a nova instância classificada de acordo com os três vizinhos mais próximos (K=3). Neste caso, o K-NN classificou a nova instância na classe positivo, visto que utiliza como critério o fato da maioria dos seus vizinhos pertencer à classe positivo.

Tabela 3.1: Conjunto de treinamento KNN.

Atributo 1	Atributo 2	Classe
0.5	0.7	positivo (+)
0.9	1.5	positivo (+)
3.8	4	negativo (-)
4.1	4.5	negativo (-)
4.4	2.5	negativo (-)
4.5	5.3	positivo (+)
6.3	7	negativo (-)
7.6	6.4	positivo (+)
7.9	6	positivo (+)

Tabela 3.2: Nova instância a ser classificada.

Atributo 1	Atributo 2	Classe
8.2	7	?

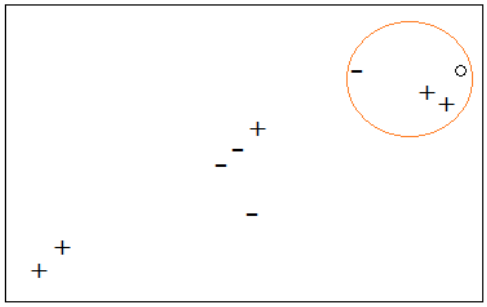

Figura 3.3: Representação de dados no espaço bidimensional

3.2.3 Redes Neurais Artificiais (*Artificial Neural Networks*)

Rede Neural Artificial (RNA) foi desenvolvida utilizando estudos do sistema nervoso biológico. Trata-se de uma técnica que constrói um modelo matemático inspirado em um sistema neural biológico simplificado. Em outras palavras, pode-se definir como um modelo computacional capaz de simular a estrutura e o funcionamento do cérebro humano, simulando a capacidade de aprendizado, associação, abstração e generalização.

A únidade básica de uma RNA é o neurônio. Ele é composto por: dendritos, os quais são responsáveis por receber estímulos transmitidos pelos neurônios; soma, que coleta e combina as informações recebidas; e axônio, que tem como função transmitir estímulos para outros neurônios. Quando um axônio entra em contato com o dendrito de outro neurônio, ou seja, quando há comunicação entre os neurônios, dá-se o nome de sinapse.

No modelo do neurônio artificial da Figura 3.4 cada terminal de entrada do neurônio recebe um valor (peso). Esses valores são ponderados e combinados usando uma função de ativação F_a para produzir uma saída.

Uma rede neural, quando utilizada para a classificação, é tipicamente um conjunto de unidades de entrada e saída (neurônios artificiais) dispostas em uma ou mais camadas e interligadas por um grande número de conexões com um peso associado (HAN; KAMBER; PEI, 2011) (Figura 3.5). O processo de aprendizagem da rede se dá por meio de um processo iterativo de ajuste

dos pesos sinápticos, de modo que ao término seja capaz de prever a classe.

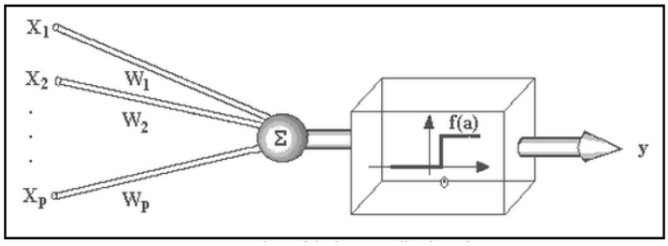

Figura 3.4: Neurônio artificial
Fonte: (MCCULLOCH; PITTS, 1943)

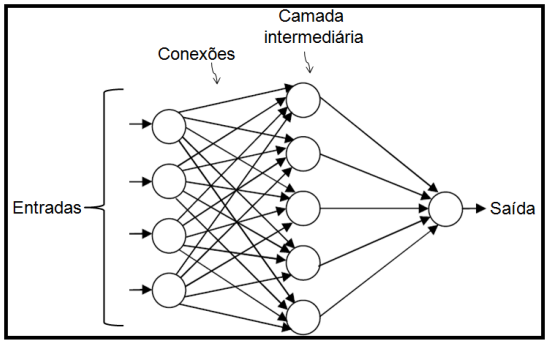

Figura 3.5: Rede neural artificial

Existem inúmeros algoritmos de RNA como: *Self-Organizing Maps (SOM)* (KOHONEN, 1990), Perceptron (ROSENBLATT, 1958), *Adaptive Linear Neuron (Adaline)* (WIDROW; STERNS, 1960) e Backpropagation (MCCLELLAND; RUMELHART, 1986). O que os diferenciam são suas especificidades (tipo de conexão entre os neurônios, número de camadas de neurônios e o tipo de treinamento utilizado, por exemplo).

3.3 MINERAÇÃO DE IMAGENS

Diferente da mineração convencional em que os dados estão representados no banco de dados na forma tabular, relacional ou de grafos, a mineração de imagens é um processo mais complexo, envolve técnicas de processamento de imagens, extração de características, indexação de objetos e visualização (RIBEIRO, 2008).

A mineração de imagens "cuida da extração de conhecimento implícito, relacionamento de imagens e outros padrões não explicitamente armazenados em bases de dados de imagens" (ZHANG; HSU; LEE, 2002).

O processo de mineração de imagens segundo Zhang et al. (2002) é dividido em quatro etapas sequenciais:

- Pré-processamento - tem como objetivo aumentar a qualidade das imagens que passarão pelo processo, realçando as características importantes e eliminando as indesejáveis;
- Transformação/extração de características - as características importantes devem ser extraídas das imagens e transformadas para um formato válido para a mineração;
- Mineração - técnicas são aplicadas com a finalidade de descobrir informações a partir dos dados;
- Interpretação/avaliação - os padrões obtidos são interpretados e avaliados por um especialista de maneira que gere conhecimento válido para a tomada de decisão.

De acordo com Silva (2006), os principais aspectos que diferem a mineração de imagens da mineração realizada em bancos de dados convencionais, são:

- Textura - os elementos das imagens (pixels) não podem ser tratados individualmente, para que não se perca a capacidade de capturar a informação de textura presente no contexto de aplicação, isto porque a análise da imagem é feita pela vizinhança de pixels;
- Processamento em vários níveis - o processamento visual acontece por partes. Inicialmente são detectadas as bordas, geometrias e estruturas dos objetos. Em seguida, os objetos são identificados na cena e contextualizados. Por fim, os elementos perceptuais, como borda, geometria e estrutura, são associados a padrões, protótipos e eventos de acordo com a percepção de cada indivíduo;
- Ambiguidade de interpretação - a interpretação está relacionada aos métodos de análise empregados, com o nível de conhecimento e experiência de quem está observando a imagem. A ambiguidade acontece quando uma imagem recebe diferentes interpretações dos observadores;
- Dependência de domínio - uma mesma imagem pode possuir diferentes informações relacionadas a diferentes domínios. A identificação dos elementos, classes e relacionamentos deve ser feita considerando o contexto em que estão inseridos.

3.4 PROCESSAMENTO DIGITAL DE IMAGENS

Desde seu surgimento na década de 60, a área de processamento digital de imagens (PDI) tem apresentado um rápido crescimento, principalmente pelo fato da maioria das imagens na atualidade serem do formato digital e dependerem de algum elemento digital de processamento.

O PDI viabiliza aplicações em duas categorias distintas: disposição de procedimentos para melhorar o aspecto visual da imagem e facilitar a interpretação humana, dentre eles, a redução de ruídos obtidos durante o processo de aquisição de imagem, restauração, realce do contraste e nitidez; processamento de dados de imagens para armazenamento, transmissão e representação, subsidiando a percepção automática de objetos de interesse por computador (GONZALEZ; WOODS, 2010).

Na Figura 3.6 encontram-se as etapas constituintes do PDI, as quais serão discutidas nas próximas subseções.

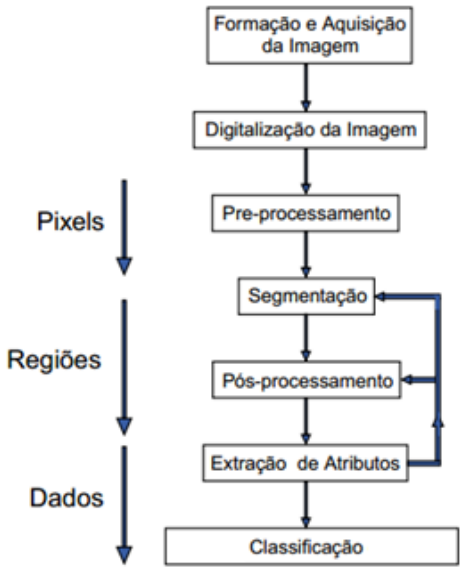

Figura 3.6: Etapas do processamento digital de imagens
Fonte: Adaptado de (ALBUQUERQUE et al., 2004)

3.4.1 Formação e Aquisição de Imagens Digitais

Como pode ser observado na Figura 3.7, "a maioria das imagens são geradas pela combinação de uma fonte de iluminação (Figura 3.7(a)) e a reflexão ou absorção de energia dessa fonte pelos objetos de uma cena 3D (Figura 3.7(b)) na qual a imagem está sendo gerada" (GONZALEZ;

WOODS, 2010).

Essa interação entre a energia irradiada e os objetos físicos é captada por sensores (Figura 3.7(c)) os quais projetam tal energia em um plano de imagem (Figura 3.7(d)). O sensor é um dispositivo físico sensível a uma faixa de energia (iluminação) no espectro eletromagnético, podendo ser raios-X, ultravioleta, espectro visível ou raios infravermelhos, que produz como saída um sinal elétrico de acordo com nível de energia detectado (MARQUES FILHO; VIEIRA NETO, 1999).

A saída da maioria dos sensores é apresentada na forma de onda de tensão contínua cuja amplitude e o comportamento no espaço trata-se do fenômeno físico a ser captado pelo sensor do sistema de imageamento. Para formar a imagem digital é necessário converter esse sinal análogo (contínuo) para um formato digital (Figura 3.7(e)), por meio das técnicas de amostragem e quantização (discutidas na Subseção 3.4.2).

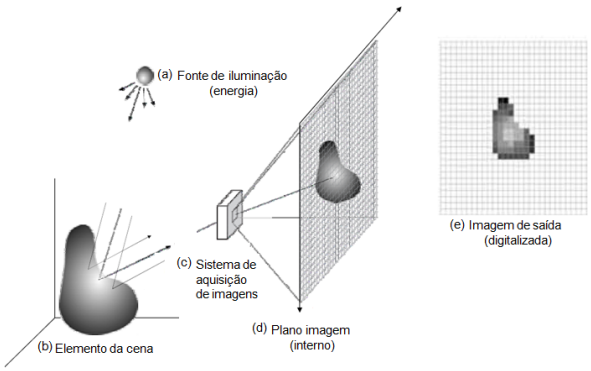

Figura 3.7: Processo de aquisição de uma imagem digital
Fonte: Adaptado de (GONZALEZ; WOODS, 2010)

Portanto, uma imagem digital pode ser definida por uma função bidimensional *f(x,y)*, da intensidade de luz refletida ou emitida por uma cena (Figura 3.8). Onde x e y são as coordenadas espaciais, e a amplitude de f em qualquer par de coordenadas é chamada de intensidade ou de nível de cinza da imagem (GONZALEZ; WOODS, 2010).

Em outra definição, imagem digital é um arquivo que pode ser manipulado e armazenado em um computador, composto por um cabeçalho contendo determinadas informações e representada por uma matriz de números que indicam a cor ou intensidade do pixel (*picture cell*) na imagem (GOMES, 2001). O pixel é o menor componente de uma imagem digital, que por sua vez é composta por um conjunto de pixels. Quanto maior o número de pixels, maior o volume

de informação armazenada, consequentemente, melhor será a qualidade da imagem.

Figura 3.8: Representação de uma imagem digital
Fonte: (MARQUES FILHO; VIEIRA NETO, 1999)

3.4.2 Digitalização da Imagem

O sinal analógico obtido à saída do dispositivo de aquisição deve ser submetido a um processo de digitalização (também chamado de discretização) para tornar seu formato apropriado para o processamento computacional. Esse processo de digitalização é dividido em duas etapas: amostragem e quantização.

Na Figura 3.9(a) é possível observar um exemplo de imagem contínua *f*, a qual será submetida à digitalização dos valores de coordenada *x,y* (amostragem) e de amplitude (quantização) para transformar-se em formato digital.

No gráfico da Figura 3.9(b), os valores de amplitude (nível de intensidade) da Figura 3.9(a) são representados ao longo das retas *AB*. O processo de amostragem se dá pela coleta de um conjunto de amostras ao longo da linha *AB* para representar o sinal, ou seja, a imagem original.

Na Figura 3.9(c) a posição das amostras no gráfico são indicadas por um sinal vertical abaixo da linha *AB* e representadas por quadrados brancos na função. A quantização dos níveis de intensidade se dá pela atribuição de um dos oito valores da escala (que varia do branco até o preto) à cada amostra. Por fim, na Figura 3.9(d) é mostrado o resultado da amostragem e quantização da imagem da Figura 3.9(a).

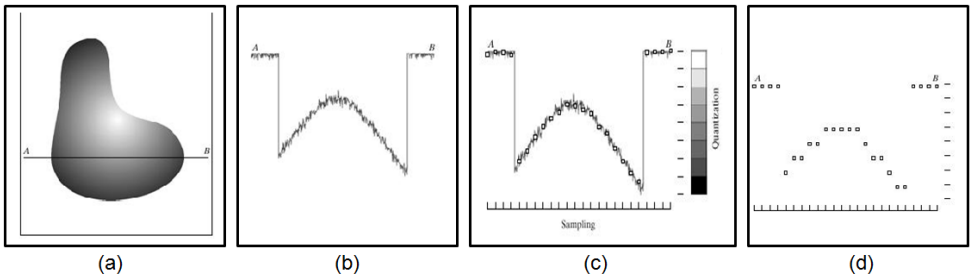

Figura 3.9: Conversão de uma imagem contínua para o formato digital
Fonte: Adaptado de (GONZALEZ; WOODS, 2010)

Na amostragem, a imagem analógica também pode ser representada por uma matriz de *M* x *N*, onde cada ponto é chamado de pixel (MARQUES FILHO; VIEIRA NETO, 1999):

$$f(x,y) = \begin{bmatrix} f(0,0) & f(0,1) & ... & f(0,N-1) \\ f(1,0) & f(1,1) & ... & f(1,N-1) \\ \vdots & \vdots & \vdots & \vdots \\ f(M-1,0) & f(M-1,1) & ... & f(M-1,N-1) \end{bmatrix}$$

Na quantização cada pixel assume um valor inteiro, na faixa de 0 a $2^n - 1$. Esse valor indica a quantidade de níveis de cinza representados por cada pixel da imagem. Assim, quanto maior o valor de *n*, maior será o nível de cinza presente na imagem digitalizada que consequentemente apresentará melhor resolução.

3.4.3 Pré-processamento de Imagens Digitais

Operações de pré-processamento podem se referir à restauração e realce de imagens, visando corrigir defeitos provenientes das variações na iluminação da cena, condições atmosféricas e de ruídos do sensor, por exemplo. Os ruídos são ocasionados devido à irregularidades ou erros que ocorrem durante a aquisição da imagem, gravação ou transmissão de dados, os quais influenciam diretamente na qualidade da imagem.

Para melhorar a visualização e interpretação das imagens por parte do observador existem inúmeras técnicas de realce, sua escolha e utilização está diretamente relacionada ao contexto e a preferência humana no que se refere a uma "imagem boa" (GONZALEZ; WOODS, 2010).

As técnicas de restauração também são direcionadas à melhora do aspecto visual da imagem, com a diferença de que sua escolha não é subjetiva e são baseadas em modelos matemáticos ou probabilísticos de degradação de imagens (GONZALEZ; WOODS, 2010).

Em geral, na etapa de pré-processamento são realizadas operações para manipular a imagem de modo que o resultado obtido seja mais satisfatório que a imagem original. Dentre as operações pode-se citar manipulação do brilho, expansão do contraste, correção de iluminação irregular, redução de ruído e realce de bordas (GOMES, 2001). Neste trabalho serão aplicadas técnicas de equalização de histograma, transformação de negativo e limiarização como operações de pré-processamento para evidenciar regiões que apresentam lesões em mamografias.

Equalização do Histograma

O histograma corresponde a um gráfico que mostra a freqüência de pixels com o mesmo tom de cinza na imagem (GONZALEZ; WOODS, 2010). É base para diversas técnicas de processamento, inclusive para o realce das imagens. Ao observar o histograma é possível obter informações de características que influenciam na qualidade da imagem como a intensidade média e espalhamento dos valores de tons de cinza, e o contraste da imagem.

O histograma normalizado é dado pela fórmula da Equação 3.4:

$$P_r(r_k) = \frac{n_k}{n} \tag{3.4}$$

Onde:

- $0 \leq r_k \leq 1$;
- $k = 0, 1, ..., L-1$, onde L é o número de níveis de cinza da imagem;
- n é número total de pixels da imagem;
- n_k trata-se do número de pixels cujo nível de cinza corresponde a k;
- $P_r(r_k)$ é a probabilidade do K-ésimo nível de cinza.

Na Figura 3.10 é possível observar exemplos de imagens e seus respectivos histogramas de acordo com os níveis de cinza apresentados. Nota-se que no histograma da Figura 3.10(a) os elementos estão concentrados no lado esquerdo (escuro) da escala de intensidades, correspondendo a uma imagem predominantemente escura. No histograma da Figura 3.10(b) acontece o oposto, os pixels estão concentrados na parte cinza da escala, caracterizando uma imagem

clara. No histograma da Figura 3.10(c) a concentração dos pixels está nos valores intermediários de cinza, resultando em uma imagem de brilho médio. No histograma da Figura 3.10(d) os elementos encontram-se bem distribuídos ao longo de toda a escala, o que significa que a imagem apresenta bom contraste. Por fim, na Figura 3.10(e), tem-se um histograma com duas concentrações de pixels, nos valores escuros e nos valores claros da escala, representando uma imagem de alto contraste.

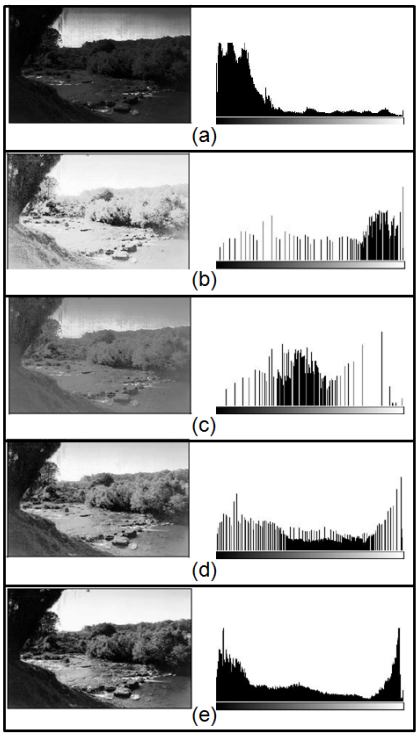

Figura 3.10: Exemplos de histogramas
Fonte: Adaptado de (MARQUES FILHO; VIEIRA NETO, 1999)

Existem várias técnicas de manipulação de histograma, uma das mais usuais é a equalização. Essa técnica visa a redistribuição dos valores de tons de cinza dos pixels de uma imagem para obter um histograma uniforme, de modo que a frequência de pixels nos níveis de cinza seja a mesma (MARQUES FILHO; VIEIRA NETO, 1999).

A maneira mais usual de equalizar um histograma é por meio da função de distribuição acumulada (*Cumulative Distribution Function (CDF)*). Essa função calcula para cada tom de cinza na imagem de entrada, um tom de cinza na imagem de saída. A mesma pode ser expressa pela Equação 3.5:

$$S_k = T(r_k) = \sum_{j=0}^{k} \frac{n_j}{n} = \sum_{j=0}^{k} p_r(r_j) \tag{3.5}$$

onde:

- $0 \leq r_k \leq 1$;
- $k = 0, 1, ..., L-1$;
- n_j é o número de pixels cujo nível de cinza corresponde a j na imagem de entrada;
- n é o número total de pixels da imagem de entrada;
- $P_r(r_j)$ é a função de distribuição de probabilidade de tons de cinza j que define o histograma da imagem de entrada.

Transformação de Negativo

Dentre as técnicas de realce de imagem encontra-se a transformação de negativo. A aplicação dessa técnica permite a reversão dos níveis de intensidade de uma imagem, ou seja, os tons de cinza escuros da imagem de entrada são transformados para tons de cinza claros na imagem de saída. Na Figura 3.11(a) tem-se um exemplo de mamografia apresentando uma lesão, a qual é mais evidênciada na Figura 3.11(b) após a aplicação da técnica do negativo. Essa técnica é indicada para realçar detalhes brancos ou cinza em regiões escuras de uma imagem (GONZALEZ; WOODS, 2010).

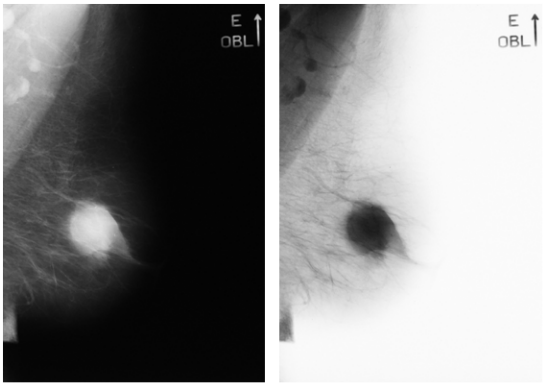

Figura 3.11: Imagem: (a) original (b) negativo

O negativo de uma imagem digital é obtido através da fórmula da Equação 3.6:

$$S = L - 1 - r \tag{3.6}$$

Onde:

- $k = 0, 1, ..., L-1$, onde L é o número de níveis de cinza da imagem;
- r é o tom de cinza do pixel a ser negativado;
- S é a imagem de saída.

Limiarização

Limiarização (*thresholding*) é uma técnica muito utilizada no contexto de segmentação de imagens por ser um dos métodos de processamento de imagens mais simples de implementação e por apresentar velocidade computacional (GONZALEZ; WOODS, 2010).

A técnica baseia-se na partição do histograma da imagem para converter todos os pixels cujo tom de cinza é maior ou igual a um certo valor de limiar T em brancos e os demais em pretos. Devido essa característica, essa técnica também recebe o nome de binarização.

A limiarização recebe como entrada uma imagem $f(x,y)$ com N níveis de cinza e produz como saída uma imagem $g(x,y)$ cujo número de níveis de cinza é menor que N. Normalmente, $g(x,y)$ apresenta dois níveis de cinza, onde 1 corresponde ao objeto e 0 ao fundo da imagem, e T é um valor de tom de cinza predefinido denominado limiar que é estabelecido de acordo com a Equação 3.7:

$$g(x,y) = \begin{cases} 1 & se f(x,y) > T \\ 0 & se f(x,y) \leqslant T \end{cases} \tag{3.7}$$

Em casos que se tem uma imagem apresentando dois tipos de objetos claros (a,b) sob um fundo escuro(c), tem-se uma limiarização multipla (Equação 3.8):

$$g(x,y) = \begin{cases} a & se f(x,y) > T_2 \\ b & se T_1 < f(x,y) \leq T_2 \\ c & se f(x,y) \leq T_1 \end{cases} \tag{3.8}$$

Quando o valor de T é uma constante aplicável a imagem inteira, tem-se uma limiarização global. Caso o valor de T se altere ao longo da imagem, tem-se uma limiarização variável.

Neste trabalho, a limiarização é aplicada como operador de pré-processamento em uma

etapa que antecede a segmentação. Na Figura 3.12(a) tem-se uma imagem de mamografia apresentando uma lesão benigna. E na Figura 3.12(b) tem-se o resultado da aplicação da técnica de limiarização. Neste exemplo foi definido o limiar T igual a 100 para a identificação da região contendo a lesão na mamografia.

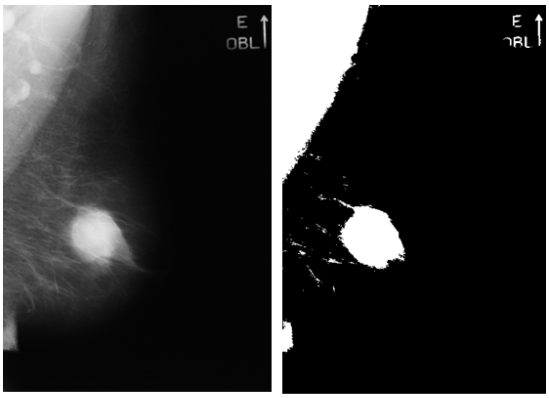

Figura 3.12: Imagem: (a) original (b) limiarizada

3.4.4 Segmentação de Imagens

A segmentação consiste na subdivisão de uma imagem em suas regiões constituintes até que os objetos de interesse sejam identificados (GONZALEZ; WOODS, 2010). Esse processo traduz para o computador o processo cognitivo realizado pela visão humana de separar objetos contidos em uma imagem baseando-se na proximidade, similaridade e continuidade dos aspectos captados. Portanto, é considerada como uma das tarefas mais complexas do processamento de imagem (NEVES; PELAES, 2001), visto que uma segmentação imprecisa influencia na etapa de extração de atributos e até mesmo na análise e classificação ao término do PDI.

Existem várias técnicas de segmentação de imagens, a maioria é baseada em propriedades de valores de intensidade como (GONZALEZ; WOODS, 2010). Estas são:

- Descontinuidade: a imagem é segmentada baseando-se em mudanças bruscas de intensidade. Essas variações formam os contornos dos objetos presentes nas imagens por meio de pontos isolados, linhas, segmentos ou curvas (MARQUES FILHO; VIEIRA NETO, 1999). Nesta categoria são incluídos os algoritmos de detecção de bordas;

- Similaridade: a imagem é dividida em regiões que se assemelham baseando-se em determinados critérios preestabelecidos como o tom de cinza ou textura. As técnicas que

utilizam-se desse fundamento são limiarização, crescimento de regiões (*region growing*) e *Watershed* os quais serão abordados neste trabalho.

Crescimento de Regiões

Uma imagem é composta por um conjunto de regiões, que por sua vez são formadas por um conjunto de pixels. Com isso, a segmentação baseada em crescimento de regiões (*Region Growing (RG)*) consiste no agrupamento das regiões de acordo com a similaridade dos tons de cinza dos pixels, cor ou textura (SANTOS, 2002). A definição dos critérios de similaridade dependem do problema abordado e do tipo de imagem a ser trabalhada. No caso de imagens de sensoriamento remoto, por exemplo, a análise é realizada pela cor. Por outro lado, em imagens de radiografia, como a mamografia, analisa-se os níveis de intensidade dos pixels.

O algoritmo de segmentação por crescimento de regiões requer dois parâmetros: um valor de inicial de similaridade e um valor que representa a área mínima da região vizinha. O mesmo funciona da seguinte maneira (BINS; FONSECA; ERTHAL, 1996):

1. A imagem é segmentada em células atômicas de um ou poucos pixels, e cada pixel é considerado como uma região distinta;

2. Cada região segmentada é comparada com as regiões vizinhas para determinar se existe similaridade. O critério de similaridade é dado por um teste de hipótese que testa a média entre as regiões;

3. No caso das regiões serem similares elas são unificadas e o nível de cinza médio da nova região é atualizado;

4. Segue a comparação com todos os vizinhos até que não haja mais regiões passíveis de unificação, assim, a região para de crescer e é rotulada como uma região completa;

5. O processo é realizado em todas as regiões incompletas, repetindo toda a sequência até que todas as regiões estejam rotuladas.

Na Figura 3.13 pode-se observar uma imagem segmentada com o algoritmo de crescimento de regiões.

Figura 3.13: Imagem: (a) original (b) segmentada

Watershed

Watershed (bacias hidrográficas) trata-se de uma técnica que se baseia em conceitos de topografia e hidrologia para a segmentação de regiões. Originou-se do princípio que uma imagem em níveis de cinza pode ser vista como um relevo topográfico, onde as regiões mais escuras são representados por vales e as regiões mais claras por montanhas. Supondo que ocorra uma inundação, seja por água vinda de cima, como se fosse chuva ou vinda de baixo, como se um orifício estivesse sido perfurado, a água vai escorrer e ocupar as partes mais baixas do terreno. Quando a água acumulada nas diversas *watersheds* está prestes a se juntar, barragens/represas ou linhas divisoras de águas chamadas *watersheds* são construídas para impossibilitar a fusão (GONZALEZ; WOODS, 2010). A inundação acaba quando apenas os topos das barragens são visíveis acima da linha da água.

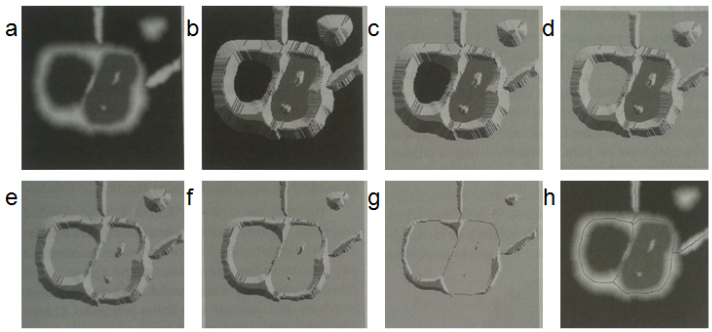

Figura 3.14: (a) Imagem original (b) Imagem topográfica (c) a (e) Diferentes fases de inundação (f) Fusão da água (g) Barragens maiores (h) Linhas da watershed

Fonte: (GONZALEZ; WOODS, 2010)

Na Figura 3.14(a) é mostrada uma imagem em níveis de cinza. Já na Figura 3.14(b) pode-se visualizar o relevo topográfico da imagem. Nas Figuras 3.14(c), (d) e (e) são apresentadas

imagens em diferentes fases de inundação. Na Figura 3.14(f) é mostrado o inicio da fusão de água entre duas *watersheds*. Na Figura 3.14(g) nota-se que foi construída uma barragem maior entre duas *watersheds* evitando que as águas dessas regiões se juntassem com águas do fundo. Na Figura 3.14(h) tem-se o resultado da aplicação da técnica, onde é possível visualizar as linhas das *watersheds*. A produção desses contornos fechados para delinear o limite dos objetos da imagem é a principal vantagem da utilização desta técnica.

As principais definições de transformada *watershed* são: inundação (*Flooding*-WT), distância topográfica (TD-WT), condição local (LC-WT), *image foresting transform* (IFT-WT), zona de empate da IFT-WT e *watershed cut* (WC-WT) (KORBES; LOTUFO, 2010).

Neste trabalho é utilizada a definição de *watersheds* por IFT, a qual baseia-se na teoria dos grafos para representar uma imagem. Nesta abordagem, os vértices (nós) do grafo correspondem aos pixels da imagem. As arestas que ligam os vértices recebem um peso por meio de uma função de dissimilaridade que fornece o valor máximo da intensidade do nível de cinza entre os nós u e v (KLAVA, 2009).

A definição IFT-WT faz uso de marcadores (pontos específicos na imagem por onde deve começar o alagamento), os quais podem ser internos (associados ao objeto de interesse) e externos (associados ao fundo da imagem) (GONZALEZ; WOODS, 2010). Cria-se uma floresta na qual existe um caminho de custo mínimo entre o conjunto de marcadores e cada vértice do grafo (pixel). Ao término do processo, obtem-se uma floresta, a qual é composta por árvores. Assim, cada conjunto de árvores cujas raízes tenham o mesmo rótulo, corresponde a uma região de interesse da imagem (KLAVA, 2009).

3.4.5 Pós-processamento de Imagens Digitais

O intuito da etapa de pós-processamento de imagens consiste na correção de possíveis erros e na eliminação de ruídos obtidos durante a etapa de segmentação. A solução para problemas como esses seria a utilização de técnicas de Morfologia Matemática (MM), as quais podem ser aplicadas em diversas tarefas de processamento de imagens tais como restauração, realce, filtragem, segmentação, extração de medidas, entre outras.

A MM trata do estudo das estruturas geométricas presentes em uma imagem (MARQUES FILHO; VIEIRA NETO, 1999). E tem como objetivo a extração de informações sobre a geometria

de uma imagem desconhecida, através da transformação de outra imagem bem definida por meio de um elemento estruturante (tamanho e forma) (SOUZA; SANTOS, 2004). Na Figura 3.15 é possível visualizar as formas de alguns elementos estruturantes. Na forma do elemento cruz, por exemplo, são considerados como vizinhos do pixel central somente os 4 pixels adjacentes a ele, lateral e verticalmente.

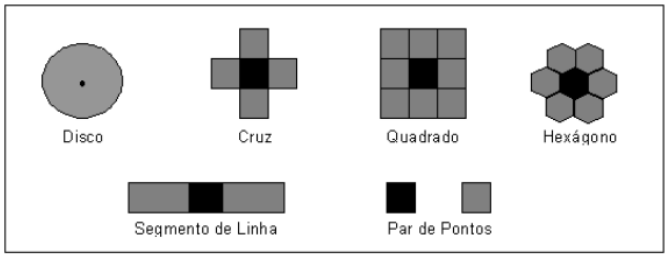

Figura 3.15: Formas de elementos estruturantes
Fonte: (SOILLE, 1999)

Por meio da MM podem ser realizadas operações para expandir os objetos da imagem (dilatação) e para diminuí-los (erosão) utilizando o elemento estruturante (vizinhança). Em outras palavras, na erosão os pixels que não atendem a um determinado padrão são apagados da imagem e na dilatação áreas relacionadas a um pixel são alteradas para um dado padrão.

Dilatação

Sejam A e B conjuntos no espaço Z^2 e seja ϕ o conjunto vazio. A dilatação de A por B, denotada $A \bigoplus B$, é definida pela Equação 3.9:

$$A \oplus B = \left\{ x \mid \left(\hat{B}\right)_x \cap A \neq \phi \right\} \tag{3.9}$$

O processo consiste em obter a reflexão de B sobre sua origem e em seguida deslocar esta reflexão de x. Assim, a dilatação de A por B é o conjunto de todos os x deslocamentos para os quais a interseção de $\left(\hat{B}\right)_x$ e A inclui pelo menos um elemento diferente de zero.

Erosão

Sejam A e B conjuntos no espaço Z^2. A erosão de A por B, denotada $A\Theta B$, é definida pela Equação 3.10:

$$A\Theta B = \{x \mid (B)_x \subseteq A\} \tag{3.10}$$

A erosão de A por B resulta no conjunto de pontos x tais que B, transladado de x, está contido em A.

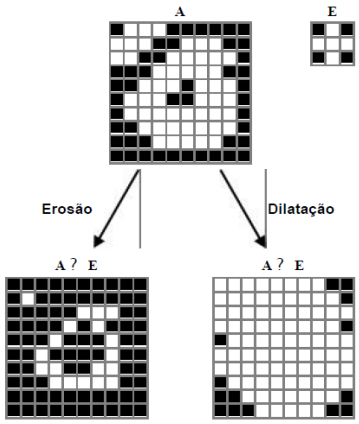

Figura 3.16: Exemplo de erosão e dilatação
Fonte: (GOMES, 2001)

Na Figura 3.16 é possível observar o efeito da erosão e da dilatação de uma imagem (A), de 10x10 pixels, por meio do elemento estruturante (E), de 3x3 pixels, onde na erosão cada pixel branco na imagem de entrada que tem ao menos um pixel branco vizinho foi invertido na imagem de saída, já na dilatação, cada pixel preto na imagem de entrada com pelo menos um pixel branco vizinho foi invertido na imagem de saída.

3.4.6 Extração de Atributos

A extração de atributos é a etapa do processamento digital de imagens que consiste na obtenção de atributos relevantes para representar e diferenciar os tipos de objetos presentes na imagem e posteriormente analisá-los.

Existem inúmeros atributos para representar regiões de interesse (ROIs) em imagens médicas, eles são distribuídos nas principais categorias: textura e geometria (morfologia).

Intuitivamente, textura descreve características da superficíe de um objeto na imagem, tais

como rugosidade, uniformidade e regularidade. De acordo com Gonzalez e Woods (2010), as três abordagens mais utilizadas no PDI para descrever a textura de uma região são:

- Estatísticas: produzem caracterizações da textura como suave, rugosa, granulada, etc;
- Estruturais: descrevem a textura como arranjos de primitivas de imagem. A primitiva é um elemento fixo que se repete numa área da imagem, como linhas paralelas espaçadas regularmente;
- Espectrais: baseadas em propriedades do espectro de Fourier, é possível detectar características como periodicidade global em uma imagem mediante a identificação de picos de alta energia no espectro.

O reconhecimento de características de textura é um desafio, visto que não apresentam um padrão regular e são dependentes de escala (FERREIRA; BORGES, 2005). Além disso, a análise de textura geralmente é um processo muito demorado, por isso não é indicado em situações onde um grande volume de imagens está envolvido (COSTA; HUMPIRE-MAMANI; TRAINA, 2012).

Em relação aos atributos geométricos, eles descrevem as propriedades morfológicas das regiões de interesse como área, perímetro e circularidade (AL-SHAMLAN; EL-ZAART, 2010). Possuem alta representatividade para identificar objetos de interesse em imagens (AL-SHAMLAN; EL-ZAART, 2010) e com isso axiliar no diagnóstico médico.

3.4.7 Classificação

Na última etapa do PDI, classificação, são aplicadas técnicas de aprendizagem e reconhecimento de padrões para transformar os dados obtidos na etapa de extração de atributos em informações úteis e relevantes.

No contexto de processamento de imagens, o objetivo da classificação é a identificação de características ou padrões das estruturas encontradas em uma imagem e de suas atribuições à determinadas classes (SOLOMON; BRECKON, 2011). Em casos em que se tem conhecimento a priori sobre os atributos e suas respectivas classes, a classificação é supervisionada, caso contrário, a classificação é não supervisionada e os atributos são agrupados por critérios de similaridades formando assim os *clusters*.

De maneira geral, a classificação têm a finalidade de substituir a análise visual dos dados por técnicas de análise automática para a identificação de objetos na imagem (QUEIROZ, 2003). A técnica de classificação, assim como, os métodos classificadores são explorados na Subseção 3.1.2 e Seção 3.2, respectivamente.

3.5 CLASSIFICAÇÃO DE TUMORES DE MAMA: TRABALHOS RELACIONADOS

A literatura abrange várias metodologias de esquemas CAD voltados ao diagnóstico do câncer de mama. Dentre elas pode-se citar o trabalho de Asad et al. (2011) que teve como objetivo a classificação de tumores de mama nas categorias maligno e benigno. Para isso utilizou um conjunto de 33 mamografias obtidas a partir do banco de dados *Mammographic Image Analysis Society (MIAS)*, as quais foram segmentadas pelo método de limiarização local (*local thresholding*) para separar a ROI da imagem. Foram extraídos e classificados 7 atributos geométricos pelo algoritmo de redes neurais *Kohnan Neural Networks*.

Em Surendiran e Vadivel (2012) foram classificadas mamografias utilizando atributos extraídos da forma geométrica e da margem da ROI. No total foram extraídos 17 atributos de 940 imagens da base de dados *Digital Database for Screening Mammography (DDSM)*. Para a identificação das ROIs nas mamografias foi utilizado o algoritmo de limiarização. Na classificação foi utilizado o CART, algoritmo baseado em árvore de regressão.

Já em Radovic et al. (2013) foram utilizadas 322 imagens do repositório MIAS, as quais passaram pela fase de pré-processamento para a aplicação da técnica de limiarização local para a remoção do músculo peitoral e identificação da região de interesse. Foram extraídos 20 atributos de textura com o objetivo de classificar mamografias como normal (sem tumor) ou anormal (com tumor). Os classificadores utilizados foram o *Support Vector Machine (SVM)*, *Naive Bayes*, *K-Nearest Neighbor*, *Logistic Regression*, Árvore de Decisão (C4.5), *Random Forest* e *Multilayer Perceptron*.

Por fim, o trabalho Liu e Tang (2014) teve como foco métodos de seleção de atributos para a classificação de tumores de mama. Foram utilizadas 826 imagens do repositório DDSM e extraídas 31 características de forma e textura. Os autores desenvolveram o método *Spatial Fuzzy C-Means Clustering* que provê a melhor inicialização para a segmentação da ROI. Este método é baseado no *level set*, o qual é originado do modelo de Contornos Ativos (*Snakes*), para

a identificação das bordas das regiões. Também foram investigados vários métodos de seleção de atributos, o mais eficiente selecionou 12 características, as quais foram classificadas pelo Support Vector Machine (SVM) e KNN.

3.6 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foram descritos os principais conceitos do processo de DCBD, mineração de dados e processamento digital de imagens, os quais dispõem de métodos e técnicas aproriadas para o desenvolvimento deste trabalho. No próximo capítulo será apresentada a metodologia e o sistema proposto.

CAPÍTULO 4

METODOLOGIA E MAMOCAD

Neste capítulo encontram-se descritas a metodologia proposta, as funcionalidades do MAMOCAD, sistema de auxílio ao diagnóstico do câncer de mama, e as tecnologias empregadas para o desenvolvimento do mesmo.

4.1 METODOLOGIA PROPOSTA

A metodologia proposta neste trabalho encontra-se dividida em duas etapas: treinamento e classificação. Na etapa de treinamento são realizadas as seguintes tarefas (Figura 4.1):

- Seleção: a mamografia que será submetida ao processo é selecionada no banco de imagens. Na etapa de treinamento as imagens selecionadas já possuem um diagnóstico;
- Transformação: a imagem é submetida a uma transformação. É aplicado um operador de pré-processamento (negativo, equalização de histograma ou limiarização) para evidenciar as lesões;
- Segmentação: o algoritmo de processamento digital de imagem (*region growing* ou *watershed*) é aplicado para segmentar a região de interesse;
- Extração de atributos: da ROI são obtidos atributos como área, perímetro, perímetro-área, shape, fractal, circularidade, entre outras;
- Criação do modelo: atributos extraídos a partir de um conjunto de ROIs de imagens e seus respectivos diagnósticos são utilizados como entrada pelo algoritmo classificador (J48, CART, MLP ou IBK);
- Modelo: caracteriza os tipos de lesões em mamografias que passaram pela etapa de treinamento. O modelo é utilizado na etapa de classificação.

A etapa de classificação é semelhante à etapa de treinamento:

- Seleção: a mamografia que passará pelo processo de classificação é selecionada. Nesta etapa a mamografia não possui diagnóstico;
- Transformação: é escolhido um operador para ser aplicado às imagens visando o realce da lesão;
- Segmentação: o objetivo desta tarefa consiste na identificação da ROI e segmentação da mesma;
- Extração de atributos: são obtidos atributos (shape, circle, fractal, etc.) da ROI;
- Classificação: os atributos extraídos são submetidos ao modelo gerado na etapa anterior (treinamento) para que este processo de classificação indique o diagnóstico da lesão;
- ROI classificada: tem-se como resultado dessa etapa a região de interesse classificada como benigna ou maligna. Esta informação pode auxiliar o médico fornecendo uma segunda opinião sobre o caso ou aumentando a certeza no diagnóstico.

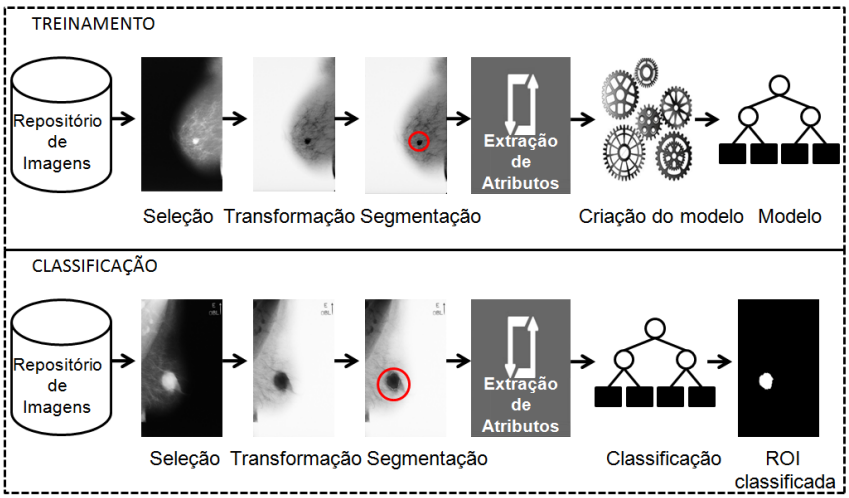

Figura 4.1: Visão geral da metodologia

4.2 MAMOCAD

O *Computer-Aided Diagnosis in Mammography (MAMOCAD)* trata-se de um sistema CAD com a finalidade de dar suporte à decisões de radiologistas ou médicos especialistas no diagnóstico do câncer de mama. A tela inicial conta com as seguintes funcionalidades (Figura 4.2):

- Configurar banco de dados: permite realizar a configuração de acesso do software ao banco de dados;
- Registrar paciente: cadastro de um novo paciente no sistema com seus respectivos dados;
- Procurar imagens: possibilita a busca por imagens já diagnosticadas armazenadas no banco de dados de acordo com as especificações (identificador, sexo e idade do paciente, localização do tumor, densidade mamária e classificação/diagnóstico) desejadas;
- Diagnosticar: a imagem referente ao paciente cadastrado é submetida ao processo para a realização do diagnóstico.

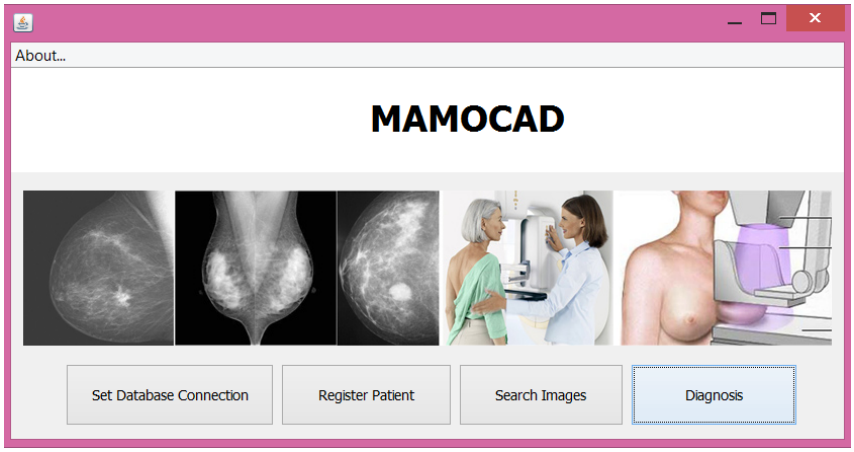

Figura 4.2: Tela inicial MAMOCAD

Caso o usuário deseje realizar um novo diagnóstico, deve-se selecionar uma imagem e escolher dentre os operadores de pré-processamento (transformação de negativo, equalização de histograma ou limiarização), o mais adequado para melhorar a qualidade da imagem e visualização das lesões. Na Figura 4.3 é possível visualizar um exemplo de mamografia e as diferentes aplições de operadores de pré-processamento. Neste caso específico, os operadores que melhor realçaram a ROI na imagem foram transformação de negativo e limiarização.

A segmentação pode ser realizada pelo *region growing* ou pelo *watershed*, ambos tratam-se de algoritmos de segmentação local semiautomática, realizando a segmentação apenas da região indicada pelo usuário. Na Figura 4.4(a) pode-se visualizar a ROI segmentada pelo *region growing* após a transformação de negativo, os respectivos atributos extraídos, bem como a classe de treinamento (maligno) selecionada. Para a classificação pode-se optar entre o J48, CART,

IBK e o Multilayer Perceptron (MLP). Na Figura 4.4(b) tem-se o resultado da classificação de uma ROI como maligna, utilizando o algoritmo de mineração J48.

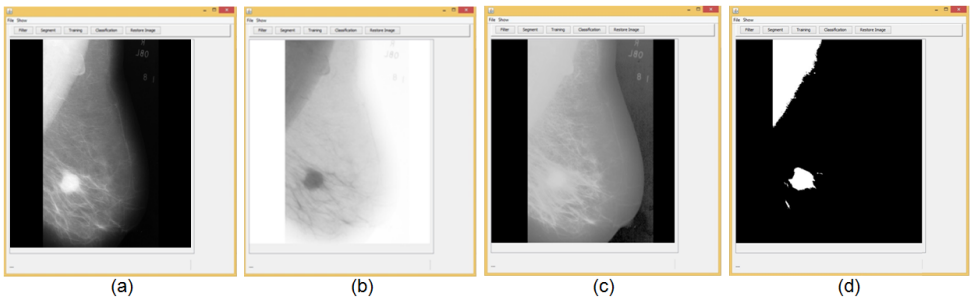

Figura 4.3: (a) Imagem original (b) Transformação de negativo (c) Equalização do histograma (d) Limiarização

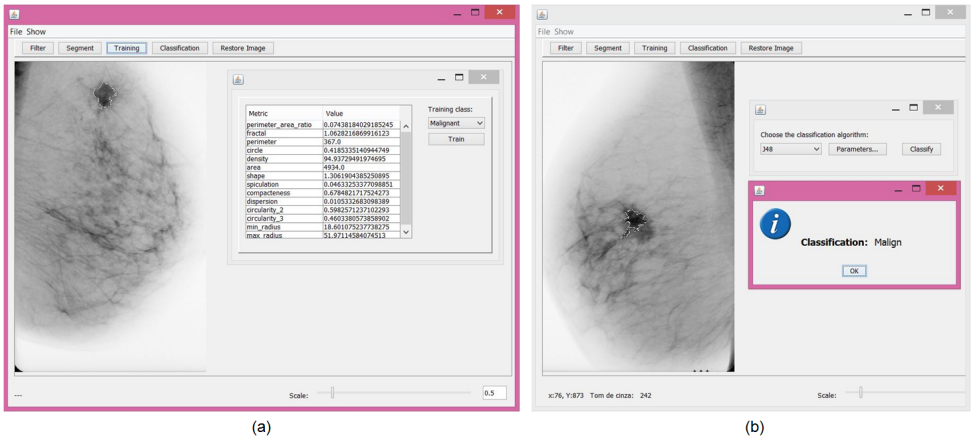

Figura 4.4: (a) Etapa de treinamento (b) Etapa de classificação

4.2.1 Tecnologias Utilizadas

Visando o desenvolvimento do MAMOCAD foram utilizadas as seguintes tecnologias:

- Java - linguagem de programação orientada a objetos desenvolvida na década de 90 pela Sun Microsystems. Utilizada neste trabalho para o desenvolvimento do MAMOCAD com a finalidade de prover portabilidade (programa que pode ser desenvolvido e utilizado independentemente do sistema operacional) ao sistema e facilitar a integração com o WEKA;

- *Waikato Enviroment for Knowledge Analysis (WEKA)* - programa de código aberto desenvolvido em Java, na Universidade de Waikato, Nova Zelândia, bastante popular no meio acadêmico.

O pacote WEKA inclui diversos algoritmos de aprendizagem de máquina para a execução de tarefas de mineração de dados como, por exemplo, classificação, agrupamento, regressão, e associação (WEKA, 2014). A Interface de Programação de Aplicações (API) do WEKA foi integrada ao MAMOCAD para a utilização dos algoritmos J48 (C4.5), CART, IBK (KNN) e MLP;

- NetBeans - ambiente de desenvolvimento integrado (do inglês *Integrated Development Environment (IDE)*), o qual fornece aos desenvolvedores ferramentas necessárias para criar aplicativos para desktop, Web e dispositivos móveis. Suporta a utilização das linguagens Java, HTML5, PHP, C, C++, entre outras. Pode ser executado nos sistemas operacionais Windows, Linux, Solaris e MacOS;

- PostgreSQL - sistema gerenciador de banco de dados objeto-relacional que por ser um dos SGBDs (sistema gerenciador de banco de dados) de código aberto mais avançados na atualidade o torna muito utilizado. Neste trabalho, ele foi utilizado para manipular os dados extraídos das mamografias, assim como para armazenar as imagens segmentadas e laudadas;

- *Java Advanced Imaging (JAI)* - trata-se de uma API desenvolvida pela *Sun Microsystems* que fornece métodos de manipulação, visualização e processamento de imagens digitais. Utilizado no contexto deste trabalho para o processamento de imagens de mamografias.

4.2.2 Processo de Desenvolvimento do MAMOCAD

Neste trabalho foi adotado o processo incremental de desenvolvimento de software. Nesse modelo os requisitos são definidos, implementados, validados e entregues incrementalmente de acordo com sua importância e a necessidade do cliente (SOMMERVILLE, 2011). Uma das vantagens de sua utilização consiste em proporcionar aos clientes uma maior flexibilidade na definição dos requisitos, podendo ser adiado o detalhamento dos mesmos. Uma vez que um subconjunto de funcionalidades é entregue, o cliente pode utilizá-lo e detalhar com mais propriedade os requisitos do próximo incremento. Isso garante que o software seja desenvolvido de acordo com as especificações e necessidades do cliente. Além disso, é natural que a cada integração de incrementos, as primeiras funcionalidades sejam mais testadas e aprimoradas, o

que diminui o número de erros do sistema.

No MAMOCAD inicialmente foram implementados e testados os algoritmos de segmentação, objetivando exatidão na delimitação das ROIs. Em seguida, identificou-se a necessidade da aplicação de operadores de pré-processamento para facilitar a segmentação das imagens. Em seguida, os atributos geométricos foram definidos, implementados e testados até se obter um conjunto capaz de representar as lesões nas mamografias. Os algoritmos de classificação J48, CART, MLP e IBK foram implementados. E paralelamente ao desenvolvimento dessas funcionalidades, se deu a implementação da interface gráfica, objetivando maior usabilidade para facilitar a interação do usuário com o sistema.

Na Figura 4.5 encontram-se ilustradas as fases do processo de desenvolvimento incremental. Ao término do desenvolvimento de um incremento, o mesmo é validado e entregue. O processo se repete integrando os incrementos até a obtenção do software completo.

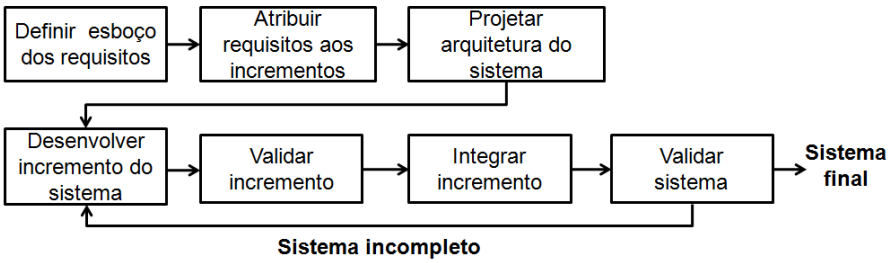

Figura 4.5: Processo de desenvolvimento incremental
Fonte: (SOMMERVILLE, 2011)

Em relação ao padrão de projeto seguido neste trabalho tem-se o *Model View Control (MVC)*. Como pode ser visualizado na Figura 4.6, este padrão consiste na divisão da arquitetura do sistema em três camadas físicas, a saber (GAMMA; HELM; JOGNSON, 2000):

- Visão (*view*) - camada de interface, a qual o usuário interage realizando ações/fornecendo dados, e que apresenta os resultados do processamento;
- Controlador (*controller*) - camada intermediária em que os comandos executados pelo usuário na camada de visão são repassados para esta e processados, modificando o modelo;
- Modelo (*model*) - camada que fornece acesso aos dados armazenados repassados pelo

usuário. Trata-se da lógica de negócio, em que os dados são gerenciados (armazenados, manipulados e gerados).

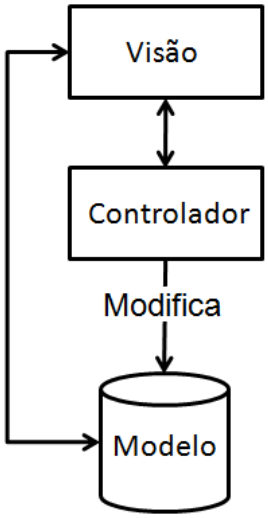

Figura 4.6: Visão geral do padrão MVC

A divisão do software em camadas apresenta como principais vantagens a flexibilidade e reuso (GAMMA; HELM; JOGNSON, 2000) do código da interface ou do controlador sem que o modelo seja alterado.

4.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foram descritas as etapas da metodologia que deu origem ao MAMOCAD, bem como as funcionalidades do software, as tecnologias utilizadas, o processo de desenvolvimento de software e a arquitetura MVC. No Apêndice A encontra-se a documentação do MAMOCAD, a qual inclui o diagrama de Caso de Uso e o diagrama de Classes. No próximo capítulo será apresentado o estudo de caso realizado para avaliar o MAMOCAD.

CAPÍTULO 5

ESTUDO DE CASO

Neste capítulo será apresentado o banco de imagens utilizado para a realização do estudo de caso, assim como a descrição dos experimentos para a avaliação do MAMOCAD, os resultados obtidos e a discussão dos mesmos.

5.1 *BREAST CANCER DIGITAL REPOSITORY* - BCDR

O BCDR (*Breast Cancer Digital Repository* - Repositório Digital de Câncer de Mama) é resultado de uma colaboração entre CETA-CIEMAT, o Instituto INEGI do Porto, a Faculdade de Medicina da Universidade do Porto e a Universidade de Aveiro. Atualmente ele disponibiliza dois repositórios de domínio público: (1) contendo filmes de mamografias digitalizadas, composto por 1010 (998 mulheres e 12 homens) casos de pacientes com idade entre 20 e 90 anos de idade; e (2) composto por mamografias digitais, com 237 casos (BCDR, 2012).

Para cada caso existe em média 4 imagens de mamografias, dentre elas, as incidências mamográficas mediolateral oblíqua (MLO) e crânio-caudal (CC) das mamas esquerda e direita. Caso a imagem apresente alguma anomalia, a mesma também é disponibilizada com a segmentação manual da lesão realizada por radiologistas especializados na área.

As imagens encontram-se classificadas de acordo com o sistema BI-RADS em uma das 9 categorias (listadas na Subseção 2.4.3). Além disso, também é informado o quadrante da mama que contém a lesão, o tipo (nódulo, microcalcificações, calcificações ou distorção de arquitetura, entre outros), idade e sexo do paciente, o resultado da biópsia (maligno, benigno, insuficiente ou suspeito) e a composição de tecido glandular na mama.

5.2 EXPERIMENTOS

Neste trabalho foram realizados experimentos para definir o conjunto de atributos geométricos, avaliar os operadores de pré-processamento, bem como os algoritmos de segmentação e o desempenho dos classificadores.

Como citado anteriormente, no MAMOCAD encontram-se os algoritmos de mineração J48, CART, IBK e MLP, todos de classificação supervisionada. O J48 foi incluido nesta metodologia pois foi utilizado por Silva et al. (2005) em uma outra abordagem, na qual se obteve sucesso na distinção de diferentes padrões espaciais presentes em imagens de sensoriamento remoto, utilizando atributos tais como área, perímetro e shape.

Visando comparar os resultados do J48, o CART foi implementado, uma vez que se trata de um algoritmo de árvore de regressão, apropriado para a aplicações com atributos contínuos. Assim como o J48 e o CART, o multilayer perceptron (MLP) e o IBK foram escolhidos devido à ampla utilização em trabalhos da área. Perez, Guevara e Silva (2012), Surendiran e Vadivel (2012), Radovic et al. (2013) e Braz Junior et al. (2006) utilizaram tais métodos em suas abordagens, o que garante a aplicabilidade dos algoritmos. Também baseando-se na aplicabilidade foi definida a utilização dos segmentadores (*region growing* e *watershed*) e operadores de pré-processamento (transformação de negativo, equalização de histograma e limiarização).

Para a realização dos experimentos, os parâmentros dos algoritmos de classificação foram alterados e ajustados objetivando o aumento da taxa de acerto. Para testar a capacidade de generalização dos classificadores, aplicou-se o método de validação cruzada (*cross-validation* - 10 *folds*), que consiste em dividir o conjunto de dados em 10 subconjuntos do mesmo tamanho, realizar 10 iterações, cada uma utilizando um subconjunto diferente para teste e os demais para treinamento.

O repositório de mamografias digitalizadas do BCDR foi o escolhido para a realização de experimentos neste trabalho. Essa escolha se deu devido ao desafio, visto que as mamografias convencionais têm qualidade inferior comparada à mamografias digitais, o que dificulta a visualização das lesões pelo especialista. Acredita-se que obtendo sucesso nos experimentos com imagens digitalizadas, a metodologia desenvolvida pode ser aplicada com êxito nas demais imagens de mamografia, inclusive, nas originalmente digitais.

Foram utilizadas apenas mamografias com incidência MLO, pelo fato de prover a visualização da mama completa e do músculo peitoral, apresentando nódulo (tumor) maligno ou benigno. Ou seja, foram descartadas dos testes imagens com diagnóstico inconclusivo, normal (com ausência de lesões) ou apresentando outros tipos de anomalias.

5.2.1 Definindo o conjunto de atributos geométricos

Tumores de mama são normalmente distinguíveis visualmente pela sua forma e margem na mamografia. Uma vez que tumores benignos tendem a possuir formato mais arredondado/ovalado e margem bem definida, e os tumores malignos possuírem forma e margem irregular (Subseção 2.4.2). Decidiu-se trabalhar com a categoria de atributos geométricos por focar em características como forma e margem da região.

Por meio de pesquisas na literatura foi definido empiricamente um conjunto de atributos, os quais foram extraídos de imagens de mamografias e submetidos à classificação pelo algoritmo J48. Uma vez que a árvore de decisão seleciona os atributos mais relevantes que contribuíram para o aumento do ganho de informação, foi definido o conjunto de 12 atributos mais representativos de regiões de interesse em imagens de mamografia (SILVA, 2006) (SURENDIRAN; VADIVEL, 2012) (TODD; NAGHDY, 2004). Tais atributos podem ser observados na Tabela 5.1, estes foram extraídos das imagens e utilizados em todos os experimentos deste trabalho.

Tabela 5.1: Atributos geométricos.

Atributo	Equação	Descrição	Aplicação
Area (A)		Retorna a área da região. Quando medida em pixels é igual ao número destes.	Tumores malignos são maiores que tumores benignos.
Perimeter (P)		Retorna o perímetro da região. Quando medida em pixels é igual ao número destes na borda da região.	Usado para calcular atributos que refletem a complexidade e compactação da região.
Fractal	$2\frac{log(0.25*P)}{log(A)}$	Índice que mede a complexidade da forma da região.	A forma dos tumores benignos é menos complexa comparada à forma dos tumores malignos.

Max radius		Retorna a distância máxima entre o centro e a borda da região.	O raio máximo das regiões de tumores malignos é maior que o raio máximo de regiões de tumores benignos com a mesma área.
Min radius		Retorna a distância mínima entre o centro e a borda da região.	Usado para calcular o atributo que reflete a circularidade da região.
Circle	$1 - \frac{A}{\pi(raio^2)}$	Retorna 0 para regiões circulares e próximo de 1 para regiões lineares.	Tumores benignos tem forma circular/oval.
Circularity	$\sqrt{\frac{MinRadius}{MaxRadius}}$	Mede a similaridade da região com elipse.	Tumores benignos podem ter forma oval.
Compactness	$\left(\frac{2\sqrt{A\pi}}{P}\right)$	Retorna o grau de dissimilaridade entre a região e um círculo perfeito.	A forma de tumores malignos diferem de um círculo perfeito.
Dispersion	$\frac{MaxRadius}{Area}$	Mede a irregularidade de uma região.	Tumores malignos tem forma irregular.
Shape	$\frac{P}{4\sqrt{A}}$	Retorna 1 para regiões compactas e aumenta de acordo com a irregularidade.	Tumores benignos apresentam forma mais compacta que regiões de tumores malignos.
Perimeter-Area	$\frac{P}{A}$	Razão entre o perímetro e a área da região. É um indicador de complexidade de forma da região.	Tumores malignos tem forma irregular.

Spiculation	$Si = \frac{l_i}{b_i^2}$	Razão entre o comprimento da borda da região e o quadrado da largura da região. Onde L é o comprimento da borda da região, b é o comprimento da base da região.	Identifica tumores malignos com forma espiculada.

5.2.2 Experimento 1

- Objetivo: avaliar técnicas de segmentação.

Para este experimento foram utilizadas imagens em seu formato original, ou seja, sem a aplicação de operadores de pré-processamento ou quaisquer alterações. Foram selecionadas 40 mamografias (20 malignas e 20 benignas), as quais foram segmentadas pelos algoritmos *region growing* e *watershed*, extraídos os atributos das ROIs e submetidos ao treinamento. Para a classificação utilizou-se 20 mamografias (10 malignas e 10 benignas) que passaram pelo mesmo processo de segmentação da ROI e extração de atributos.

Region growing

A segmentação realizada pelo *region growing* se deu de maneira local e semiautomática, onde o operador indica a região suspeita de conter a lesão na imagem, fornece a distância euclidiana e o algoritmo realiza o processo. O parâmetro da distância euclidiana foi alterado de acordo com as especificidades de cada imagem até que a segmentação abrangesse corretamente toda a região de interesse.

Tabela 5.2: Resultado da classificação após a segmentação com *region growing*.

Region growing.			
Algoritmo	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
J48	90	80	85
CART	90	80	85
MLP	70	90	80
IBK	70	90	80

Na Tabela 5.2 encontram-se os resultados da classificação com os algoritmos J48, CART, MLP (taxa de aprendizagem = 0.3 e 4 camadas de neurônios) e IBK (K=3)

Watershed

Assim como o *region growing*, o *watershed* realiza a segmentação local de maneira semiautomática. O processo ocorre de acordo com o parâmetro *grayscale* (nível de cinza dos pixels) informado pelo usuário, o qual deve ser ajustado de acordo com as particularidades de cada imagem e lesão.

Os resultados de classificação dos algoritmos J48, CART, MLP (taxa de aprendizagem = 0.4 e 2 camadas de neurônios) e IBK (K=3) encontram-se na Tabela 5.3.

Tabela 5.3: Resultado da classificação após a segmentação com *watershed*.

Watershed			
Algoritmo	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
J48	90	70	80
CART	70	70	70
MLP	70	70	70
IBK	70	80	75

5.2.3 Experimento 2

- Objetivo: avaliar operadores de pré-processamento.

Neste experimento utilizou-se o mesmo conjunto de imagens do Experimento 1, onde 40 imagens passaram pelo treinamento e 20 pela etapa de classificação. Aplicou-se o operador de pré-processamento (equalização do histograma/limiarização/ transformação de negativo), em seguida, realizou-se a segmentação da ROI com o *region growing*, visto que o mesmo obteve os resultados mais satisfatórios do Experimento 1.

Equalização do histograma

Na Tabela 5.4 encontram-se os resultados da classificação das imagens com a equalização do histograma, utilizando os algoritmos J48, CART, bem como o MLP (com taxa de aprendizado = 0.3 e 2 camadas de neurônios) e o IBK (K=1). A equalização do histograma é um processo

rápido e não necessita que o usuário insira parâmetros, visto que essa técnica redistribui os valores de tons de cinza dos pixels para obter uma imagem mais uniforme e de melhor contraste.

Tabela 5.4: Resultado da classificação após a equalização do histograma.

Equalização do histograma			
Algoritmo	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
J48	100	60	80
CART	100	60	80
MLP	100	70	85
IBK	80	70	75

Limiarização

Na Tabela 5.5 encontram-se os resultados do experimento com a aplicação da limiarização no pré-processamento das imagens. Para este experimento, utilizou-se os algoritmos J48, CART, MLP (taxa de aprendizado=0.4 e 3 camadas de neurônios) e IBK (K=3). Diferente da equalização do histograma, a limiarização necessita que o usuário indique o valor do limiar *T*, esse parâmetro foi ajustado de acordo com cada imagem para segregar a ROI.

Tabela 5.5: Resultado da classificação após a limiarização.

Limiarização			
Algoritmo	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
J48	70	90	80
CART	70	90	80
MLP	90	90	90
IBK	80	80	80

Transformação de negativo

É possível visualizar na Tabela 5.6 o desempenho dos algoritmos J48, CART, MLP (taxa de aprendizado=0.4 e 7 camadas de neurônios) e IBK (K=1) na classificação de imagens com a aplicação da transformação de negativo. Essa técnica não requer do usuário a inserção de parâmetros, pois trata-se de uma técnica de reversão dos níveis de intensidade de pixels.

Tabela 5.6: Resultado da classificação após a transformação de negativo.

Transformação de negativo			
Algoritmo	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
J48	90	90	90
CART	80	90	85
MLP	90	90	90
IBK	90	70	80

5.2.4 Experimento 3

- Objetivo: avaliar o desempenho dos algoritmos de classificação.

Apesar da grande quantidade de casos existentes no repositório de imagens BCDR, o fato desta pesquisa envolver apenas a avaliação de tumores (massas/nódulos) limitou a quantidade de imagens para serem trabalhadas. Dessa forma, foram selecionadas cerca de 134 mamografias de incidência MLO esquerda e direita apresentando tumores benignos ou malignos. Desse total, 70 imagens (35 malignas e 35 benignas) foram submetidas ao processo de treinamento e 30 (15 malignas e 15 benignas) à classificação. A quantidade maior de imagens dedicadas ao treinamento se deu devido à importância de um modelo com um grande número de exemplos e de preferência, bem representativos. As demais imagens foram descartadas para não comprometerem o resultado do experimento, visto que trata-se de imagens de baixa qualidade que impossibilitaram a realização da segmentação correta das ROIs. Então, o experimento 3 se deu da seguinte maneira:

Após a seleção das imagens aplicou-se o operador de transformação de negativo, o qual contribuiu para a segmentação da ROI e aumentou as taxas de acerto dos algoritmos classificadores do Experimento 2. Utilizou-se o *region growing* para a segmentação, extraiu-se os 12 indicadores da forma geométrica das regiões para criar o modelo de treinamento e, em seguida, realizar a etapa de classificação das ROIs como benignas ou malignas pelos algoritmos J48, CART, MLP e IBK.

No caso do algoritmo MLP os resultados mais satisfatórios foram obtidos utilizando taxa de aprendizado 0.3 e 2 camadas de neurônios. A melhor configuração do IBK se deu com o valor de K igual a 7. Já os algoritmos de árvore de decisão (J48 e CART) foram utilizados sem

aplicar o método de poda. É importante ressaltar que apesar de algumas imagens conterem mais de uma lesão, para cada processo de diagnóstico apenas uma ROI é treinada/classificada. Na Tabela 5.7 encontram-se os resultados deste experimento.

Tabela 5.7: Resultado da classificação final.

Algoritmo	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
J48	100	93	96
CART	100	93	96
MLP	100	80	90
IBK	80	100	90

5.3 RESULTADOS E DISCUSSÃO

5.3.1 Experimento 1

Region growing

Na Tabela 5.2 encontram-se os resultados deste experimento. Observa-se que as taxas de classificação foram iguais para os algoritmos de árvore de decisão (J48 e CART), os mesmos apresentaram as maiores taxas de sensibilidade (90%) e de acurácia (85%). Já os algoritmos MLP e IBK obtiveram sensibilidade igual a 70%, especificidade igual a 90% e 80% de acurácia.

Observou-se que o *region growing* (RG) foi mais preciso na delimitação das lesões malignas. A explicação para isso está relacionada ao fato do algoritmo RG basear-se na intensidade do nível de cinza para o crescimento das regiões. Uma vez que as regiões de tumores malignos tem alta densidade comparada à regiões de tumores benignos, essas regiões densas possuem resposta mais alta e destacam-se na imagem, o que facilita a identificação e segmentação das mesmas. Por outro lado, determinadas lesões benignas têm nível de cinza similar a regiões de tecidos sadios, por isso, a dificuldade do RG na segmentação de tumores benignos. Outra limitação do RG está na segmentação de lesões próximas ao tecido glandular denso, visto que são obscurecidas, dificultando a segregação da ROI do fundo da imagem.

Apesar das limitações, a segmentação apresentou precisão, contribuindo para a extração de atributos que refletiram nas taxas de acerto dos classificadores, e eficiência, tornando o processo de diagnóstico produtivo e menos cansativo.

Watershed

Apesar do *watershed* ser um dos métodos de segmentação mais utilizados na literatura, o mesmo não alcançou a precisão necessária na delimitação das ROIs (malignas e benignas) do conjunto de imagens utilizado. Por se basear na similaridade dos níveis de cinza para realizar a segmentação, apresentou limitações similares ao *region growing*, tais como na segmentação de lesões benignas com nível de cinza similar a regiões de tecidos sadios e de lesões próximas ao tecido glandular denso, as quais tendem a ser obscurecidas. Acredita-se que estes tenham sido os principais fatores contribuintes para a obtenção dos resultados expostos na Tabela 5.3, onde observa-se que as taxas de classificação foram inferiores às obtidas na Tabela 5.2, em que a segmentação foi realizada pelo *region growing*.

Neste experimento realizado com o *watershed*, a maior taxa de sensibilidade foi obtida pelo J48 (90%), já a maior taxa de especificidade se deu por meio do IBK (80%). De forma geral, a acurácia dos classificadores neste experimento foi inferior à acurácia do experimento aplicando o *region growing*.

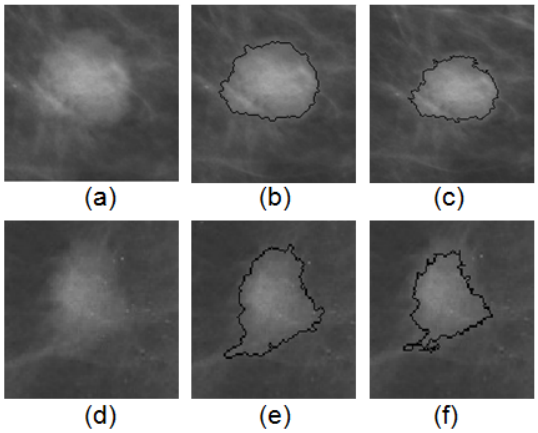

Figura 5.1: (a) e (d) ROIs (b) e (e) Segmentação *region growing* (c) e (f) Segmentação *watershed*

Na Figura 5.1 encontram-se duas regiões de interesse, as quais foram segmentadas pelo *region growing* e *watershed*. Nota-se que o contorno realizado pelo *region growing* é mais preciso.

5.3.2 Experimento 2

Equalização do histograma

Na Tabela 5.4 são mostrados os resultados da classificação em que aplicou-se como operador de pré-processamento a equalização do histograma. Ao comparar os resultados deste experimento com os da Tabela 5.2, no qual não foram aplicadas operações antecedendo a segmentação, observa-se que houve aumento nas taxas de sensibilidade de todos os classificadores, porém, observou-se a diminuição das taxas de especificidade e de acurácia.

A técnica de equalização de histograma apresentou como desvantagem sua aplicação em imagens de mamas com grande porcentagem de tecido denso, pois ao nivelar a intensidade dos pixels das regiões sadias com os pixels das regiões apresentando anomalias, a lesão é parcialmente ocultada (obscurecida), dificultando assim a segmentação.

Limiarização

Comparando os resultados da Tabela 5.5 com a Tabela 5.4 em que é realizada a equalização do histograma das imagens, verifica-se que com a aplicação da limiarização houve aumento das taxas de especificidade de todos os classificadores, mas por outro lado houve a diminuição das taxas de sensibilidade.

A limiarização é amplamente utilizada para a segmentação de regiões e como operador de pré-processamento. Na abordagem utilizada neste trabalho, observou-se que é mais indicada quando já se sabe onde a lesão está localizada e com base nisso ajustar o parâmetro do limiar. Esse ajuste pode tornar o processo mais lento e inconsistente, visto que varia de acordo com o usuário.

Tranformação de negativo

Ao aplicar a transformação de negativo, de fato houve melhora na visualização e segmentação das lesões, principalmente nas malignas que apresentaram-se bem mais destacadas nas imagens. A vantagem da utilização deste operador comparado aos demais utilizados refletiu também no desempenho dos classificadores. Neste experimento (Tabela 5.6), os algoritmos apresentaram sensibilidade igual ou superior a 80% e a maioria obteve até 90% de especificidade. De modo

geral, aplicando a transformação de negativo foram obtidas as melhores taxas de classificação do Experimento 2.

5.3.3 Experimento 3

Na Tabela 5.7 encontram-se os resultados do experimento final desta dissertação. Com exatamente 70 imagens utilizadas na etapa de treinamento e 30 para a classificação, os algoritmos J48, CART e MLP alcançaram 100% de sensibilidade. Já o algoritmo IBK obteve a maior taxa de especificidade, igual a 100%. Com base nestes resultados pode-se comprovar a eficácia do método.

A transformação de negativo contribuiu para facilitar a visualização e segmentação das ROIs. Porém, mesmo com a aplicação do operador, o RG apresentou dificuldade na segmentação de imagens de baixa qualidade, nas quais até mesmo o médico/especialista seria incapaz de detectar lesão. Sendo assim, essas imagens foram descartadas para não comprometerem o resultado final da classificação.

Observou-se que para o conjunto de imagens utilizadas, o RG demonstrou-se preciso na delimitação da ROI, o que resultou na extração de atributos representativos contribuindo para a elevada taxa de acerto de classificação. Além disso, acredita-se que o sucesso da classificação deve-se à etapa de treinamento e o número de exemplos utilizados, visto que, baseados em um modelo os classificadores são capazes de dividir corretamente as instâncias nas respectivas classes.

Na Tabela 5.8 é realizado um comparativo com os trabalhos apresentados anteriormente na Subseção 3.5. Observa-se que no trabalho de Radovic et al. (2013) utilizou-se 20 atributos de textura e foi obtido 79.33% de acerto na classificação. Em Asad et al. (2011), com apenas 7 atributos extraídos da forma geométrica das ROIs atingiu-se a taxa de 80% de acerto. Surendiran e Vadivel (2012) utilizaram o maior número de imagens para teste e 17 atributos de forma e margem, o mesmo alcançou 93.72% de acerto. Mais recentemente, Liu e Tang (2014) obtiveram 94% de acerto extraindo 12 atributos de forma e textura de 826 imagens. No método proposto neste trabalho, utilizando-se um conjunto de 12 atributos geométricos extraídos de 100 imagens foi obtido 96% de acurácia.

Com base nesta comparação, conclui-se que características da forma geométrica são mais

relevantes do que atributos extraídos da textura das ROIs. Além disso, outro fator determinante para o sucesso da classificação é a combinação dos atributos utilizados. Dessa forma, é importante maximizar a utilização de atributos representativos e minimizar o tamanho do conjunto (redução da dimensionalidade). Também observa-se a importância de utilizar, para o modelo de treinamento, imagens que representem com eficácia os tipos de lesões. Assim, acredita-se que submetendo-se mais imagens à etapa de treinamento, a taxa de falso positivo do método proposto tende a diminuir, aumentando a especificidade.

Tabela 5.8: Comparação trabalhos relacionados.

Trabalho	Classificação	Atributos	Imagens	Método	Acurácia
Asad et al. (2011)	Benigno *vs* maligno	7 (forma)	33 (MIAS)	*Kohnan Neural Networks*	80%
Surendiran e Vadivel (2012)	Benigno *vs* maligno	17 (forma e margem)	940 (DDSM)	Cart	93,72%
Radovic et al. (2013)	Normal *vs* anormal	20 (textura)	322 (MIAS)	C4.5	79,33%
Liu e Tang (2014)	Benigno *vs* maligno	12 (forma e textura)	826 (DDSM)	SVM	93%
Método proposto	Benigno *vs* maligno	12 (Forma)	100 (BCDR)	J48/CART	96%

5.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foram detalhados os experimentos realizados para a avaliação da metodologia e do MAMOCAD. Os resultatos foram apresentados e discutidos. O método proposto neste trabalho apresentou uma acurácia de 96% com 100 imagens, quando o melhor resultado conhecido foi de 93,72% com 940 imagens. Desta forma, conclui-se que com um universo menor de dados o método proposto obtém uma acurácia superior ao melhor registrado na bibliografia da área. No próximo capítulo serão apresentadas as conclusões e trabalhos futuros.

CAPÍTULO 6

CONCLUSÕES

Apesar da grande variedade de imagens médicas para o diagnóstico do câncer de mama, a mamografia é a mais eficaz na detecção de lesões em fase inicial. No entanto, a leitura e interpretação dessas imagens é um processo complexo que exige muita atenção e experiência do profissional.

Com intuito de amenizar este problema, neste trabalho foi proposta uma metodologia e desenvolvido o MAMOCAD, um sistema de diagnóstico auxiliado por computador para dar suporte a médicos e especialistas na interpretação e diferenciação de tumores de mama. O software viabiliza a aplicação de operadores de pré-processamento, a segmentação das ROIs e a extração de atributos geométricos para classificar tais lesões como malignas ou benignas.

Por meio dos experimentos realizados comprovou-se que operadores de pré-processamento podem contribuir para facilitar a visualização e segmentação de ROIs em mamografias. Dos operadores utilizados, a transformação de negativo demonstrou-se o mais adequado para imagens de radiografia, pois destaca regiões brancas ou cinzas em fundos escuros.

Na segmentação das mamografias, o *region growing* demonstrou-se mais preciso que o *watershed*. Contudo, o RG apresentou limitação ao segmentar determinadas regiões apresentando lesões benignas, devido à semelhança do nível de cinza do pixel com as regiões vizinhas e lesões malignas próximas a tecido glandular denso. Nos demais casos, o algoritmo demonstrou precisão na delimitação das ROIs e eficiência, tornando o processo de diagnóstico mais rápido.

A classificação foi realiza pelos algoritmos de mineração de dados J48, CART, IBK e MLP. Os métodos de árvore de decisão, J48 e CART, obtiveram os melhores resultados, possibilitaram a classificação correta de 96% das mamografias, obtiveram 100% de sensibilidade e 93% de especificidade, o que comprova a eficácia do método. Além disso, o MAMOCAD apresenta interface amigável, intuitiva e de fácil usabilidade, que visa aumentar a produtividade do profissional da saúde.

Com a realização deste trabalho é possível concluir que fatores como o tipo e a combinação de atributos influenciam diretamente na taxa de acerto de classificação. Os atributos geomé-

tricos utilizados são suficientes para representar e diferenciar tumores benignos e malignos em mamografias, o que valida a hipótese levantada neste trabalho que a partir de técnicas de PDI é possível desenvolver uma metodologia de extração de atributos da forma geométrica de regiões de interesse em mamografias, os quais submetidos ao processo de Mineração de Imagens são classificados de acordo com o tipo de lesão apresentada.

Como trabalho futuro pretende-se desenvolver e agregar ao MAMOCAD o método de Detecção Auxiliada por Computador (CADe) das regiões de interesse em imagens de mamografia, além de investigar e aplicar outros tipos de operadores para melhorar o contraste e representatividade das imagens, visto que a qualidade das mesmas influencia diretamente no desempenho do algoritmo de segmentação e consequentemente na classificação.

Também deseja-se realizar testes com um número maior de imagens e executar a etapa de pós-processamento, aplicando técnicas de morfologia matemática para eliminar ruídos obtidos na segmentação. Além disso, torna-se interessante classificar as lesões mais encontradas em mamografias (tumores e calcificações) de acordo com o sistema BI-RADS, uma vez que esta terminologia é amplamente utilizada por radiologistas e especialistas em todo o mundo.

REFERÊNCIAS

ACR. *Mammography*. [S.l.]: Illustrated Breast Imaging Reporting and Data System (BI-RADS). 4th ed. Reston: American College of Radiology, 2003.

ACR. *ACR BI-RADS Atlas: Mammography*. 2013. American College of Radiology.

ACR. *ACR BIRADS Atlas 5th Edition*. 2013. Disponível em: <acr.org/birads>.

AL-SHAMLAN, H.; EL-ZAART, A. Feature extraction values for breast cancer mammography images. *International Conference on Bioinformatics and Biomedical Technology*, p. 335–340, 2010.

ALBUQUERQUE, M. P. de et al. Análise de imagens e visão computacional. *V Escola do CBPF*, 2004. Disponível em: http://www.cbpf.br/ mpa/G7-marcio.pdf.

ASAD, M. et al. Early stage breast cancer detection through mammographic feature analysis. *Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on. IEEE*, p. 1–4, 2011.

BAUM, F. et al. Computeraided detection in direct digital full-field mammography: initial results. *European Radiology*, v. 12, n. 12, p. 3015–3017, 2002.

BCDR. *Breast Cancer Digital Repository*. 2012. Disponível em: <http://bcdr.inegi.up.pt/>.

BENVENISTE, A. P. A.; FERREIRA, A. H. P. G.; AGUILLAR, V. L. N. Dupla leitura no rastreamento mamográfico. *Radiologia B*, v. 39, n. 2, p. 85–89, 2006.

BINS, L.; FONSECA, L.; ERTHAL, G. S. Satellite imagery segmentation: a region growing approach. *In: VIII Brazilian Symposiumon Remote Sensing*, p. 677–680, 1996.

BOYD, N. F. et al. Mammographic densities and breast cancer risk. *Cancer Epidemiol Biomarkers & Prevention*, v. 7, p. 1133–1144, 1998.

BOZEK, J.; DELAC, K.; GRGIC, M. Computer-aided detection and diagnosis of breast abnormalities in digital mammography. *50th International Symposium ELMAR. IEEE*, v. 1, p. 45–52, 2008.

BRAZ JUNIOR, G. et al. Identificação de massas em mamografias usando textura, geometria e algoritmos de agrupamento e classificação. *VI Workshop de Informática Médica - WIM*, p. 94–104, 2006.

BREIMAN, L. et al. *Classification and Regression Trees*. [S.l.]: Wadsworth International Group, 1984.

CALAS, M. J. G. et al. Avaliação de parâmetros morfométricos calculados a partir do contorno de lesões de mama em ultrassonografias na distinção das categorias do sistema bi-rads. *Radiol Bras*, v. 44, p. 289–296, 2011.

CALAS, M. J. G.; GUTFILEN, B.; PEREIRA, W. C. de A. Cad e mamografia: por que usar esta ferramenta. *Radiologia Brasileira*, v. 45, n. 1, p. 46–52, 2012.

CHEN, D.-H. et al. The correlation analysis between breast density and cancer risk factor in breast mri images. *Biometrics and Security Technologies (ISBAST), International Symposium on Biometrics and Security Technologies*, p. 72–76, 2013.

COSTA, A. F.; HUMPIRE-MAMANI, G.; TRAINA, A. J. M. An efficient algorithm for fractal analysis of textures. *SIBGRAPI*, v. 25th Brazilian Symposium on Computer Graphics and Image Processing, p. 39–46, 2012.

DOI, K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, v. 31, n. 4-5, p. 198–211, 2007.

ELSHINAWY, M.; ABDELMAGEED, A. B. W.; CHOUIKHA, M. Effect of breast density in selecting features for normal mammogram detection. *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, p. 141–147, 2011.

ERNSTOFF, L. T. et al. Breast cancer risk factors in relation to breast density. *Cancer Causes Control*, v. 17, p. 1281–1290, 2006.

FACELI, K. et al. *Inteligência artificial: uma abordagem de aprendizado de máquina*. 1. ed. Rio de Janeiro: LTC, 2011. 394 p.

FAYYAD, U. M.; SHAPIRO, G. P.; SMYTH, P. *Advances in knowledge discovery and data mining*. [S.l.]: AAAI Press, 1996. 1-34 p.

FERREIRA, C. B. R.; BORGES, D. L. A selection strategy of a minimal subset of wavelet features in a multiresolution approach for the classification of tumors on mammograms. *V Workshop of Medical Informatics*, v. 1, 2005.

FIGUEIRA, R. N. M. et al. Fatores que influenciam o padrão radiológico de densidade das mamas. *Radiologia Brasileira*, v. 36, p. 287–291, 2003.

GAMMA, E.; HELM, R.; JOGNSON, R. *Padrões de Projetos: Soluções Reutilizáveis*. [S.l.]: Bookman Companhia Editora, 2000. 364 p.

GARUD, Y. G.; SHAHARE, N. G. Detection of microcalcifications in digital mammogram using wavelet analysis. *American Journal of Engineering Research (AJER)*, v. 2, p. 80–85, 2013.

GOLDSCHIMIDT, R.; PASSOS, E. *Data mining: um guia prático*. [S.l.]: Campus, 2005.

GOMES, O. D. F. M. *Processamento e análise de imagens aplicados à caracterização automática de materiais*. 2001. Dissertação (Mestrado) - Pontifica Universidade Católica do Rio de Janeiro.

GONZALEZ, R. C.; WOODS, R. E. *Digital image processing*. 3. ed. [S.l.]: Pearson Prentice Hall, 2010.

GRAM, I. T.; FUNKHOUSER, E.; TABAR, L. The tabár classification of mammographic parenchymal patterns. *European Journal of Radiology*, v. 24, p. 131–136, 1997.

GRAY, H. *Gray's anatomy*. 39. ed. [S.l.]: Elsevier, 2005.

GUY, C.; FFYTCHE, D. *An introduction to the Principles of Medical Imaging*. [S.l.]: Imperial College Press, 2005.

HAN, J.; KAMBER, M.; PEI, J. *Data mining: concepts and techniques*. 3. ed. [S.l.]: Morgan Kaufmann Publishers, 2011.

HART, J. M. *A Practical Guide to Infra-red Thermography for Building Surveys*. [S.l.]: BRE-Press, 1991. 98 p.

HENRIQUE NETO, G. et al. Processamento e armazenamento de imagens médicas. *3° Congresso Nacional de Iniciação Científica e 1° Congresso Internacional de Iniciação Científica. in: Anais do 3° CONIC e 1° COINT*, 2003.

INCA. Prevenção e controle de câncer. *Revista Brasileira de Cancerologia*, p. 317–332, 2002.

INCA. *Atlas de mortalidade por câncer*. Maio 2011. Disponível em: <http://mortalidade.inca.gov.br/Mortalidade/>.

INCA. *Câncer de mama*. 2012. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/mama>.

INCA. *O que é câncer?* Maio 2014. Disponível em: <http://www1.inca.gov.br/conteudo"_"view.asp?id=322>.

KERLIKOWSKE, K. et al. Longitudinal measurement of clinical mammographic breast density to improve estimation of breast cancer risk. *Journal of the National Cancer Institute*, v. 99, p. 386–395, 2007.

KIERSZENBAUM, A. L.; TRES, L. L. *Histologia e Biologia Celular: Uma introdução à patologia*. 3. ed. [S.l.]: Elsevier, 2012.

KIM, S.; YOON, S. Bi-rads features-based computer-aided diagnosis of abnormalities in mammographic images. *Information Technology Applications in Biomedicine*, v. 6th International Special Topic Conference on ITAB. IEEE, p. 235–238, 2007.

KLAVA, B. *Segmentação interativa de imagens via transformação watershed*. Dissertação (Mestrado) — Universidade de Sao Paulo, 2009. Disponível em: <http://www.ime.usp.br/ klava/dissertacao.pdf>.

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

KOPANS, D. *Imagem da mama*. 2. ed. [S.l.]: Revinter, 2000.

KORBES, A.; LOTUFO, R. de A. Análise de algoritmos da transformada watershed. *17th International Conference on Systems, Signals and Image Processing*, 2010.

KUHL, C. K. Current status of breast mr imaging. part 2. clinical applications. *Radiology*, v. 244, p. 672–691, 2007.

LEITE, G. C. et al. A utilização de técnicas de limiarização para auxílio no diagnóstico de câncer de mama. *III Encontro Nacional De Engenharia Biomecânica*, 2011.

LI, L. et al. Breast tissue density and cad cancer detection in digital mammography. *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005.*, v. 27th Annual International Conference of the. IEEE, p. 3253–3256, 2006.

LIBERMAN, L. et al. The breast imaging report and data system: positive predict value of mammographic features and final assement categories. *American Journal of Roentgenology*, v. 171, p. 35–40, 1998.

LIU, X.; TANG, J. Mass classification in mammograms using selected geometry and texture features, and a new svm-based feature selection method. *IEEE Systems Journal*, v. 8, n. 3, p. 910–920, 2014.

LIZARRAGA, I. M. et al. Review of risk factors for the development of contralateral breast cancer. *The American Journal of Surgery*, v. 206, n. 5, p. 704–708, 2013.

MARQUES FILHO, O.; VIEIRA NETO, H. *Processamento digital de imagens*. [S.l.]: Brasport, 1999.

MARQUES, P. M. de A. Diagnóstico auxiliado por computador na radiologia. *Radiologia Brasileira*, v. 34, n. 5, p. 285–293, 2001.

MCCLELLAND, J. L.; RUMELHART, D. E. *Parallel distributed processing: Explorations in the microstructure of cognition*. [S.l.]: MIT Press, 1986.

MCCUE, C. *Data mining and predictive analysis: intelligence gathering and crime analysis*. 1. ed. [S.l.]: Butterworth-Heinemann, 2007. 368 p.

MCCULLOCH, W.; PITTS, W. *A logical calculus of the ideas immanent in nervous activity*. [S.l.]: Bulletin of Mathematical Biophysics, 1943. 115-133 p.

MCKENNA, R. J. The abnormal mammogram radiographic findings, diagnostic options, pathology, and stage of cancer diagnosis. *Cancer*, v. 74, n. S1, p. 244–255, 1994.

MITCHELL, T. M. *Machine learning*. New York. United States of America.: McGraw-Hill, 1997.

MOHAMED, W. A.; KADAH, Y. M. Computer aided diagnosis of digital mammograms. *Computer Engineering & Systems*, v. 7, p. 299–303, 2007.

MOORE, K. L.; DALLEY, A. F. *Anatomia Orientada Para a Clínica*. 4. ed. [S.l.]: Guanabara Koogan, 2007.

MORTON, M. J. et al. Screening mammograms: interpretation with computer-aided detection - prospective evaluation. *Radiology*, v. 239, p. 375–383, 2006.

MOUSA, D. A. et al. What effect does mammographic breast density have on lesion detection in digital mammography? *Clinical Radiology*, p. 333–341, 2013.

NEVES, S. C. M.; PELAES, E. G. Estudo e implementação de técnicas de segmentação de imagens. *Revista Virtual de Iniciação Acadêmica da UFPA*, v. 1, n. 2, p. 1–11, 2001.

PAULINELLI, R. R.; CALAS, M. J. G.; FREITAS JUNIOR, R. de. Birads e ultra-sonografia mamária - uma análise crítica. *Femina.*, v. 35, n. 9, p. 565–572, 2007.

PEREZ, N.; GUEVARA, M. A.; SILVA, A. Evaluation of feature selection methods for breast cancer classifications. *15th International Conference on Experimental Mechanics*, p. 10, 2012.

QUEIROZ, C. J. de. *Análise de transformações geométricas para o georreferenciamento de imagens do satélite CBERS-I*. 2003. Dissertação (Mestrado) - UFRGS - CEPSRM.

QUINLAN, J. Induction of decision trees. *Machine Learning*, v. 1, p. 81–106, 1986.

QUINLAN, J. R. *C4. 5: programs for machine learning*. [S.l.]: Morgan kaufmann, 1993.

RADIOLOGY, A. C. of. Mammography. *Reston, VA: American College of Radiology*, Illustrated Breast Imaging Reporting and Data System (BI-RADS). 4th ed. Reston: American College of Radiology; 2003., 1993.

RADOVIC, M. et al. Application of data mining algorithms for mammogram classification. *Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on. IEEE*, p. 1–4, 2013.

RIBEIRO, M. X. Suporte a sistemas de auxílio ao diagnóstico e de recuperação de imagens por conteúdo usando mineração de regras de associação. Tese (Doutorado) - Instituto de Ciências Matemáticas e de Computação - ICMC-USP. 2008.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, v. 65, n. 6, p. 386–408, 1958.

RUSSELL, S. J.; NORVIG, P. *Artificial intelligence*. Second edition. [S.l.]: A Modern Approach, 2003. Prentice Hall.

SANTOS, V. T. *Segmentação de imagens mamográficas para detecção de nódulos em mamas densas*. 2002. Dissertação (Mestrado) - Escola de Engenharia de São Carlos - Universidade de São Paulo - USP.

SILVA, M. P. dos S. *Mineração de padrões de mudança em imagens de sensoriamento remoto*. Tese (Doutorado) — Instituto Nacional de Pesquisas Espaciais - INPE, 2006.

SILVA, M. P. dos S. *Mineração de dados: conceitos, aplicações e experimentos com weka*. 2008. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/erirjes/2004/004.pdf>.

SILVA, M. P. S. et al. Mining patterns of change in remote sensing image databases. *Fifth IEEE International Conference on Data Mining*, p. 362–369, 2005.

SIMOFF, S.; DJERABA, C.; ZAIANE, O. Multimedia data mining between promisses and problems. *SIGKDD Explorations*, v. 4, p. 118–121, 2002.

SKAANE, P.; ENGEDAL, K.; SKJENNALD, A. Interobserver variation in the interpretation of breast imaging. *Acta Radiol*, p. 497–502, 1997.

SOILLE, P. *Morphological image analysis: principles and applications*. [S.l.]: Berlin: Springer-Verlag, 1999. 316 p.

SOLOMON, C.; BRECKON, T. *Fundamentals of digital image processing: a practical approach with examples in matlab*. [S.l.]: Wiley-Blackwell, 2011. 344 p.

SOMMERVILLE, I. *Engenharia de Software*. 9. ed. [S.l.]: Pearsson, 2011.

SOUZA, A. I. de; SANTOS, C. A. dos. Morfologia matemática. p. 49–80, 2004.

SURENDIRAN, B.; VADIVEL, A. Mammogram mass classification using various geometric shape and margin features for early detection of breast cancer. *Int. J. Medical Engineering and Informatics*, v. 4, n. 1, p. 36–54, 2012.

TAHMASBI, A.; FATEMEHSAKI; SHOKOUHI, S. B. Classification of benign and malignant masses based on zernike moments. *ComputersinBiologyandMedicine*, v. 41, p. 726–735, 2011.

TAN, P. N.; STEINBACH, M.; KUMAR, V. *Introdução ao datamining: mineração de dados*. [S.l.]: Editora Ciência Moderna Ltda, 2009.

TODD, C. A.; NAGHDY, G. Method for breast cancer classification based solely on morphological descriptors. *Medical Imaging 2004: Image Processing*, v. 5370, p. 857–867, 2004.

VACEK, P. M.; GELLER, B. M. A prospective study of breast cancer risk using routine mammographic breast density measurements. *Cancer Epidemiology, Biomarkers & Prevention*, v. 13, p. 715–722., 2004.

WEKA. *Weka 3: Data Mining Software in Java*. 2014. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/index.html>.

WIDROW, B.; STERNS, S. D. Adaptive switching circuits. *IRE WESCON Convention Record*, p. 96–104, 1960.

WOLFE, J. N. Breast patterns as an index of risk for developing breast cancer. *Roentgen Ray Society*, p. 1130–1139, 1976.

ZHANG, J.; HSU, W.; LEE, M. L. Image mining: trends and developments. *Journal of Intelligent Information Systems*, v. 19, n. 1, p. 7 – 23, 2002.

APÊNDICE A

DOCUMENTAÇÃO MAMOCAD

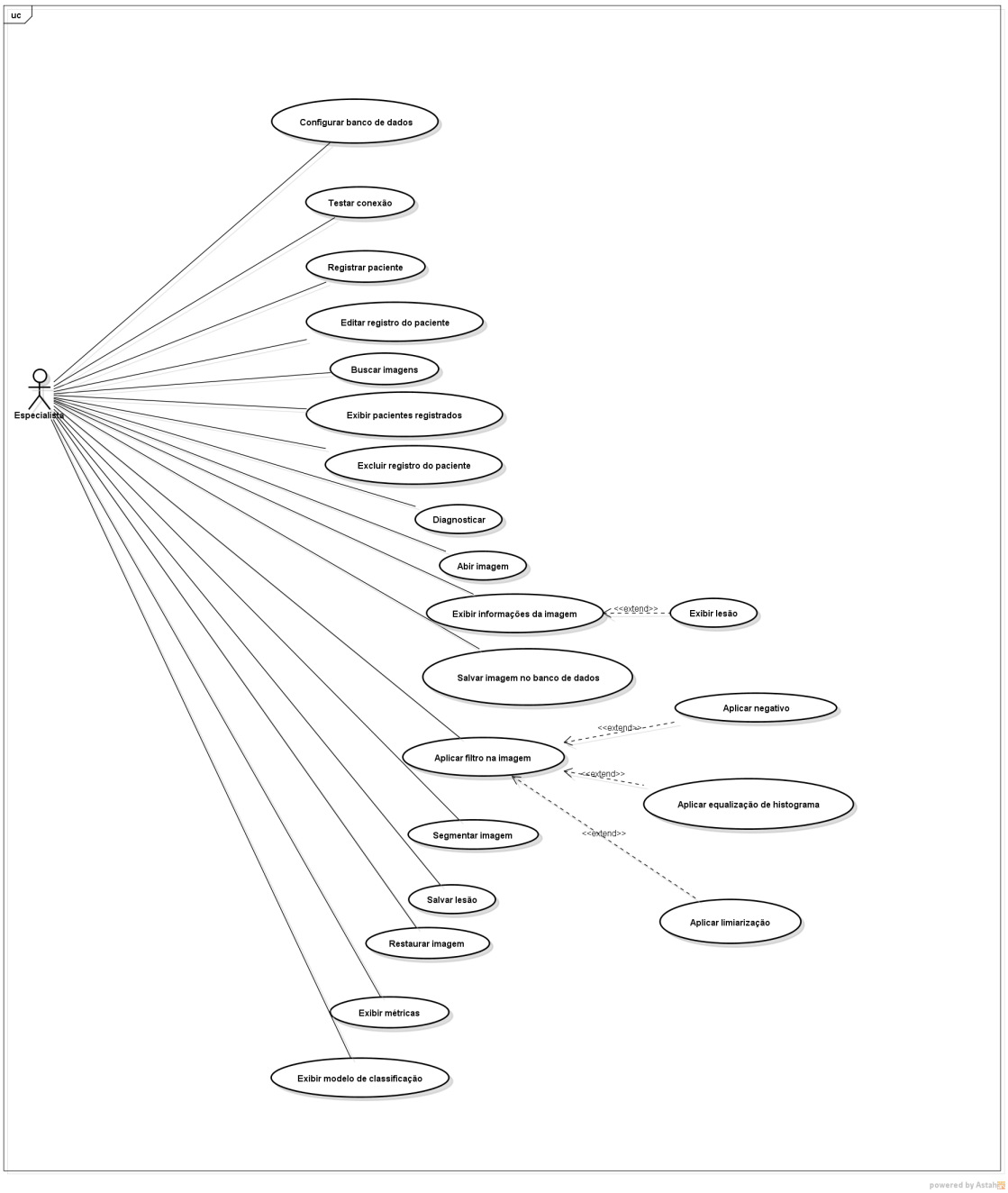

Figura A.1: Diagrama de Caso de Uso

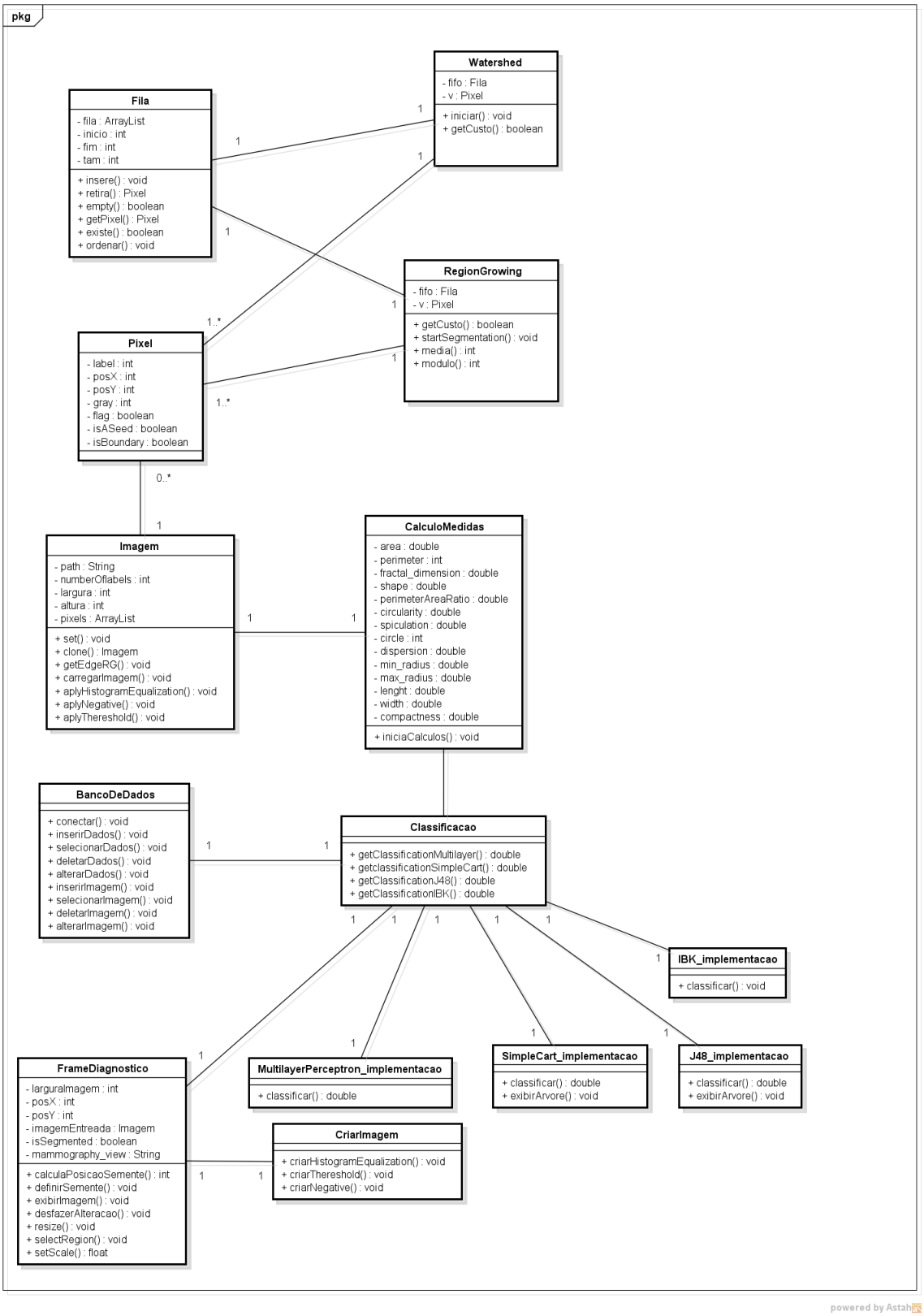

Figura A.2: Diagrama de Classes